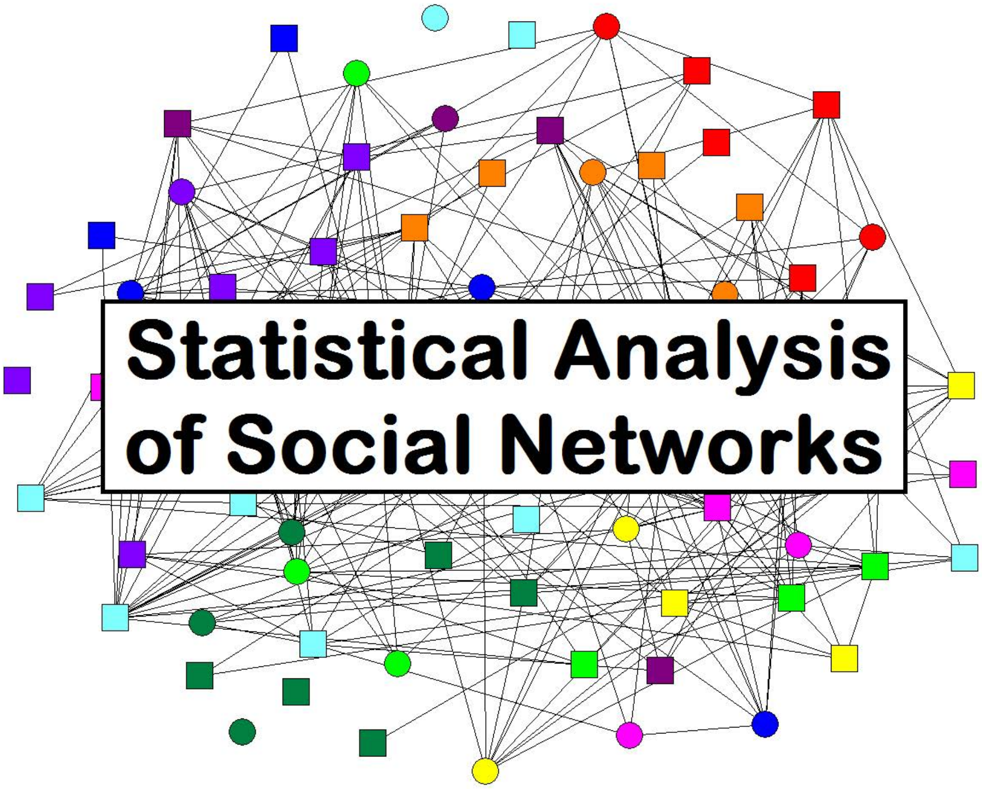


Statistical Analysis of Social Networks

An Online Open Access Textbook by Jacob Apkarian and Robert A. Hanneman

About this book

The purpose of this on-line textbook is to provide readers with an introduction to statistical applications of social network data. The text outlines the differences between variable-oriented statistical analyses and relation-oriented analyses, and using real data provided by the authors, takes the reader through examples that demonstrate how to test for association between attributes embedded in networks, between multiple networks themselves, and between the attributes and the networks they are embedded within. The text also goes on to examine models of network selection.

It is assumed that the reader has basic knowledge of statistical approaches used to describe distributions, estimate parameters of those distributions, and test hypotheses about those parameters. It is also assumed that readers have a basic understanding of social network data, and recommended that readers refer to the following text by Hanneman and Riddle (2005) when necessary: http://faculty.ucr.edu/~hanneman/nettext

Apkarian, Jacob and Robert A. Hanneman. 2016. *Statistical Analysis of Social Networks.* Jamaica, NY: City University of New York, York College (http://web.york.cuny.edu/~japkarian/).

Table of contents

Chapter 1. The Social Network Perspective

The statistical analysis of social networks is a specialized application of the general ideas of describing distributions, estimating parameters of those distributions, and testing hypotheses about those parameters. We're assuming that readers already have knowledge of these general ideas. So, to get started, let's first get an understanding of what makes the application of statistics to social network data "special."

1.1 Psycho-metrics, Econo-metrics, and Socio-metrics

At the introductory level, the applied statistics that are taught in all of the social science disciplines are pretty much the same. Students learn to describe the distributions of scores on variables measured across independently sampled cases. The notions of association, partial association, and inference from sample to population are learned. Training at the "intermediate" level is also very similar across the social sciences, where almost everyone gets a heavy dose of applications of generalizations of the general linear model for testing hypotheses about relations between variables. But, beyond this point, each of the social sciences has developed applications that are quite distinctive and attuned to needs of particular subject matter emphases.

Over-simplifying, "psycho-metrics" responds to the challenge of attempting to systematically and reliably assess latent (mental) states that cannot be measured directly. Psychometricians have developed highly sophisticated tools for working with multiple indicators, factors, and scaling. Also over-simplifying, "econo-metrics" responds to challenges of distinguishing signal from noise, characterizing trends, and assessing causal hypotheses from observational (rather than experimental) data.

"Socio-metrics" as a special flavor of formal and quantitative analysis has existed for quite some time (Moreno, 1951). Socio-metricians deal with the special problems and issues that arise when the units of analysis, across which variance is distributed, are relations between social actors, rather than attributes of individual actors.

Most graduate students being trained in quantitative analysis for Sociology learn at least the basics of the special tools of psycho-metrics and econo-metrics. Oddly, few learn anything about "socio-metrics." But, this is changing with the growing popularity of social network analysis (SNA), along with the convergence and cross-fertilization of interest in complex networks in many disciplines. The purpose of this book is to provide an introduction to thinking statistically about data that describe social relations, rather than social actors.

1.2 The Social Network Perspective

The use of graphs to represent relational data is commonplace in a wide range of sciences. The formal analysis of graphs has a very long history in mathematics and the use of statistical methods to analyze relational data has become particularly important and commonplace in physics and bio-sciences. Social scientists borrow from (and in a few cases, contribute to) these rich histories. The application of graph theory and the statistical analysis of relational data in the social sciences have a particular flavor, due to the subject matter and the theoretical questions of interest in these disciplines.

Social network analysis differs from the mainstream methodological tradition in most of the social sciences, which emphasizes the analysis of individual cases and their variable attributes (rather than relations) and experimental planned comparisons (rather than uncontrolled observational data).

Most social scientists are well versed in the more main-stream "independent-cases" and "relations-among-variables" approach to statistical analysis. Statistical methods for relational data adapt and use most of the same ideas, but with particular emphases. It's worth taking just a few minutes to get a sense of the distinctive flavor of relational, rather than variable/attribute analysis.

1.2.1 Focus on Relations

Social network analysis seeks to identify (and describe, and predict) regular patterns in the statics and dynamics of relations among social actors. The actors are most often individual humans, but they might also be populations, organizations, or symbols and cultural categories.

The emphasis on the "social," of course, is what social science is about. The emphasis on "relations" is another way of saying that what is of primary interest are "structures" composed of multiple individuals, and not the individuals themselves. Sociologists often use the adage that "sociology is not about people," by which they mean that the subject matter is regularities of social structures, not individuals.

In mainstream statistical methods, the most common approach is to examine distributions of, and associations between scores on variables, measured across individuals. Statistical methods for relational data also examine distributions of, and associations between scores – but the scores describe the relations between individuals, rather than attributes of each individual. Relational methods examine the distribution of relations, measured across pairs of individuals.

Let's suppose that we had a sample of 10 people that we were observing. For a variable-oriented analysis, the sample size is 10. We assume, for inferential purposes, that the observations are independent; or, alternatively, that we can specify the non-independence as some form of correlated error.

In a relational analysis of observations on the same 10 people, we have a "sample size" of 45 if we assume that the relations are symmetric or bonded ((10 * 9)/2), or a sample size of 90 if the relations are asymmetric or directed. That is, the unit of observation is the relation between pairs of individuals, not the individual. Obviously, these observations are not independent as multiple relations are "nested" within persons. That is, the same person (or node) is part of many of the observations.

The difference in how non-independence of cases is treated is the main technical complexity of the analysis of relational data. The kinds of statistical hypotheses, and many of the tools, are otherwise the same for variable-oriented and relation-oriented analysis.

But, there is a critical conceptual difference. Relational analysis is all about describing, testing hypotheses about, and modeling social structures, or relations between actors. Conventional variable analysis is all about describing, testing hypotheses about, and modeling relations among attributes.

1.2.2 Relations and Attributes

Social network analysis is not a substitute for attribute/variable-oriented analysis. SNA is an additional perspective that is used in conjunction with attribute/variable analysis.

For many research questions, network influences are seen as a cause or predictor of individual attributes. For example, the happiness of one's friends may influence one's own happiness. For other research questions, networks can be seen as the result of individual attributes. For example, people who are happy may be more likely to initiate friendship relations with others. As the example suggests, sometimes networks and attributes may determine one another. Individual differences may select for patterns of building networks, while individual differences may also be modified by network influences.

In addition to research questions that combine both attributes and relations, there are also some questions that may be purely relational. Do the patterns in a social network (e.g. who are friends with whom) influence patterns in another relational network of the same actors (e.g. who seeks advice from whom)? Do features of a pattern of social ties at one point in time affect the pattern of ties at a later point in time?

As we move through the chapters that follow, we will look at techniques for addressing questions where relational data are independent and dependent with attribute (variable oriented) data, as well as techniques for examining association among relational data.

1.2.3 Dynamics of and on Networks

Social network analysts commonly distinguish between "dynamics on a network" and "dynamics of a network" (or, somewhat ambiguously, "network dynamics").

Dynamics on a network assume that the relational variable(s) are fixed and affect changes in attributes. For example, the attitudes of actors who are more central in a network may be expected to be more influential on the attitudes of others than are the attitudes of actors who are more peripheral. The pattern of social relations among actors is being seen as a determinant of how the attributes of actors are related.

Dynamics of networks focus on the change in pattern of relational ties itself. The outcomes to be explained are the making and breaking of relational ties among actors. As ties are made or broken, the network changes, or becomes dynamic. Analyses of the dynamics of networks are central research questions in the broader fields of complexity and network science. In the study of social networks, the dynamics of networks is the study of how social structures change. Changes in structure may be due to inherent tendencies in structures themselves and/or due to the attributes of the actors embedded in those structures.

The statistical analysis of social networks has tool-kits for examining both dynamics on networks (usually observed as one cross-section) and tools for examining the dynamics of networks (usually observed as a time-series or fully time-continuous set of changes in relations).

1.2.4 Multiple Levels of Analysis

In variable oriented statistical analysis the individual cases or observations may sometimes be seen as “embedded” in “contexts.” This implies a degree of “non-independence” among cases which can be conceptualized as occurring within multiple levels of analysis. It is helpful to draw a distinction among three rather different ways in which multiple levels of analysis commonly enter statistical analysis of attribute/variable data.

First, sometimes cases are part of (network analysts would say “affiliated with”) larger social units. Individual students may be nested within classrooms that are nested within schools that are nested within districts or neighborhoods. Cases like this are not wholly independent observations, and mixed-models and multi-level modeling methods are often applied.

Second, cases may not be independent of other particular cases due to co-existence in some local space. In geo-spatial statistics, the attributes of a spatial area may be correlated with the attributes of adjacent spatial areas either because the boundaries of plots are arbitrary and the variables are continuously distributed in space, or because of omitted variables that have “local” influences. Spatial auto-regressive and spatial auto-correlation modeling is sometimes used when this type of non-independence exists.

Of course, the effects of adjacent cases on a focal case need not be geo-spatial; the effects may be social-spatial, or due to adjacency in a social network. Statistical methods for spatial autocorrelation and autoregression can also be applied to cases that are at known “social distances” from other cases. In the above instances, statistical corrections are used to remove non-independence due to the embedding of social actors in some higher level geographic or social space.

Third, cases might be thought of as being non-independent because they share the same or similar scores on some variable or attribute. Two persons who are both women might be thought to be non-independent because of the influence of this common attribute. This kind of non-independence is, of course, at the core of variable-oriented analysis. Here social actors can be viewed as being embedded in higher level social categories.

Social network analysis recognizes this type of non-independence in two rather different ways. If two nodes in a network share an attribute (say, both are women), network analysts would often look for "homophily" effects in their relational data. That is, the fact that two nodes have the same or a similar attribute might be hypothesized to affect the likelihood that there is a social network tie between them. Also, if there is a tie from one to the other, that it is likely to be reciprocated. Two nodes that are "closer" to one another in a network might also be likely to influence one another in the direction of becoming more similar (if the attribute in question is mutable).

Alternatively, network analysts might treat the non-independence of cases due to a common attribute as a "two-mode" network problem. We won't deal with approaches based on this way of thinking in this text, but the idea is straightforward. In the two-mode way of thinking in network analysis, cases may be one of the modes and variables the other. Associations between variables are observed when cases share the attributes. For example, if being older and being female are associated, it is because some cases that are more likely to be affiliated with the category "old" are also more likely to be affiliated with the category "female." Simultaneously, two cases are closer, or more similar, or share common affiliations if each case is tied to "old" and to "woman."

Many variables/attributes-oriented analysis questions have complexities arising from non-independence of observations, particularly in observational rather than experimental data. Many of these complexities can easily be seen as arising from the embedding of cases in networks. So, one important set of issues to be dealt with in the text that follows is how to do "conventional" or "variables and attributes" analysis in the presence of network embedding. It is useful to think of these kinds of problems as multi-level problems where cases are embedded in a network.

Many network analysis questions are also usefully cast as multi-level problems. Social networks are structures (patterns of relations among cases) that arise out of the "agency" of the individual social actors. To understand the dynamics on networks, and the dynamics of networks, the attributes of actors almost always need to be taken into account. Actors with different attributes (i.e. different scores on variables) are likely to have different networking behaviors.

The social network analysis perspective takes structures, rather than individuals, as its central concern. But, the perspective is inherently multi-level. The clearest statement of this idea continues to be that of Ronald Breiger (1974). Variables/attributes analysis takes individuals as its central concern. In many cases, though, it is also multi-level; individual cases are not independent of one another because of their embedding in social networks, contexts, or categories.

Socio-metrics, then, is a distinctive branch of quantitative analysis because it focuses on structures, or relations. But, it cannot be separated from variables/attributes analysis. Sociologists should re-cast their thinking about "conventional" statistical analysis of case-wise data to be explicit about how they treat structural effects.

1.3 Organization of the Book

Our plan for the book assumes that the reader is reasonably comfortable with conventional statistical analysis (i.e. the analysis of distributions and joint distributions of variables, measured across cases). One of our goals is to show how the analysis of variables (or "attributes" in network jargon) can be connected in powerful and useful ways with the analysis of relational data. Relational data are also used to address certain research questions and hypotheses that are unique to the social network perspective. Several of our chapters will focus on approaches and techniques for testing hypotheses about networks as the outcome to be explained.

We will begin (in Chapter 2) with a brief look at how social network data are structured. From a strictly mathematical point of view, there is nothing all that unusual about relational data – relational data are simply collections of matrices and vectors. But social network analysis does have a specialized language, and draws some analytical distinctions among types of data that are useful in helping to translate substantive problems into formal statistical analyses. There are many different software systems for working with network data, and they do vary in the details of how data are prepared for analysis (Huisman and van Duijn, 2011). Fortunately, most data structures are quite simple and it is usually easy to move data from one application to another -- which is often necessary.

We only briefly touch on descriptions of univariate distributions of variables (i.e. attributes) in Chapter 3. Descriptions of (and hypothesis tests about) univariate distributions are important, but covered in any basic courses in conventional statistics. We also won't spend much time discussing the description of and hypothesis testing about the univariate distributions of relational data. Network analysis generally, and social network analysis, particularly, have developed an extremely large number of tools for characterizing all kinds of interesting things about the shape and texture of a network. There are now a number of useful sources (including Wasserman and Faust, 1994 and Hanneman and Riddle, 2005) that cover these issues. We will spend a little time reviewing the ideas of degree-distributions and triad-censuses in our later chapters, as these are critical to understanding the theory underlying exponential random graph theory.

Chapters 3 and 4 discuss measures of association and tests of significance when dealing with "monadic" (that is attribute or "variables") data and "dyadic" (that is, network or structural level) data.

Chapter 3 takes a look at conventional attribute/variable analyses with more explicit attention to the problem of non-independence of observations that arises from network embedding. We will be looking at some approaches to understanding the association between two attributes/variables when the cases are drawn from a network rather than from independent sampling.

Chapter 4 looks at some simple approaches to studying the relationship between two networks, or the association between two dyadic or relational variables. Actors may have multiple forms of social ties that covary (e.g. both friendship and authority relations). Similarly, we may have panel data on a social relation and be interested in the correlation between earlier and later observations of the structure.

Chapters 3 and 4 look at how we study association between two attributes and between two relations, respectively. In Chapter 5, we take the next logical step by examining the association between an attribute and a relation. All three of these chapters focus on symmetric association rather than prediction and modeling of hypothesized causal relations.

In Chapter 6 we shift our focus to the study of asymmetric association, which predicts, or models hypotheses about causal effects. Chapter 6 focuses on the analysis of "network influence." That is, how are the attributes of an actor (i.e. the scores of cases on variables) affected by the ways that the node is embedded in a network, and the attributes of the "alters" to which each "ego" is connected? A very wide range of important substantive problems in sociology deal with questions of these kinds of "social influences." Do the attitudes and behaviors of those with whom I interact affect my attitudes and behaviors?

Chapter 7 turns the prediction problem around: how can we use individual's attributes to predict the ways in which they become embedded in a network? This kind of problem is often called "network selection." That is, how do the attributes of social actors shape the ways in which they make or break social relations to others – and, in the process, "select" one possible emergent network instead of another? In "network selection" problems, the relation or network is the dependent variable. This is a rather new way of thinking about things for many readers. So, Chapter 7 will spend some time looking at how SNA theorizes the processes that create networks. These ideas become quite important in understanding the remaining chapters.

In "conventional" statistics, the primary tool for problems involving the prediction and modeling where there are multiple hypothesized causal influences and need for statistical control is the generalized linear model. In Chapter 8, we tackle the same problem for networks, rather than attributes, as outcomes. At the time of the writing of this text, there are two somewhat related – but not yet fully integrated – approaches to multiple-variable prediction of networks as outcomes.

Specialists in the statistical analysis of social network data have developed a quite distinctive approach to relational variable outcomes based on the underlying theory of "exponential random graph" development. These approaches place an emphasis on using the structural tendencies of social networks (for example, the tendencies toward "reciprocity" or "closure") as predictors in explaining complex patterns of relations. The field of "exponential random graph" modeling is a distinctive approach to the analysis of relational data that is firmly grounded in social science theory, and underlies the analysis of network development and co-evolution that are discussed in Chapter 9.

The prediction of network relations as outcomes, however, can also be cast as a rather straightforward general linear mixed-model type of problem in which relations between two actors are nested in the cross of the two actors (and their attributes, as well as the attributes of the dyad). At the time of this writing, the mixed-models approach to network data has the comparative advantages of dealing with relations that are measured at the nominal, ordinal, or interval-ratio levels; exponential random graph models, to date, deal with binary outcomes. Mixed models are also familiar to many analysts, and integrate with a wide body of approaches to complicated data structures. But, so far, mixed models approaches to relational data do not have the underpinnings of SNA theories of where social structures come from and do not deal easily with the issues of structural effects and complex underlying distributions that vary with graph density – the great strength of exponential random graph models.

In Chapter 9 we take a brief look at two very important areas at the "cutting edge" (at the time of this writing) of modeling in the exponential random graph tradition. Exponential random graph theory is particularly useful as a statement of how social relations develop and change over time as actors select network structures by making and breaking social ties. Sometimes SNA data have repeated cross-sections (or "panels") of observations on the patterns of ties among the same actors as they change over time. Some models have been developed ("Sienna") specifically for studying network development of "evolution."

The earlier chapters developed two related themes. On one hand, networks develop and are shaped (i.e. one network is "selected" instead of another) by choices made by actors in forming or breaking ties. These choices may be "biased" by the attributes on the actors. On the other hand, some attributes of the actors making these choices may be influenced or shaped by the attributes or behaviors of the "alters" to which each "ego" is connected. For example, a student might experiment with drugs because his/her friends do, and consequently drop some friendship ties with non-user friends and make new ties with others who are drug users. That is, the attributes of an actor (e.g. being a drug user or not) may "co-evolve" with their position in the network (e.g. the likelihood of having friendship ties with others who use drugs). At the cutting edge of statistical applications in network analysis are some "Sienna" models that treat both actor attributes and relations as joint outcomes of joint processes of "network selection" and "network influence."

1.4 Summary

Applied statistics in the social sciences have a common set of core concepts and techniques that differ little across the disciplines. Several of the disciplines have also developed more specialized emphases that address problems that are particularly common in the types of data that arise from the research designs and measurement methods that the disciplines often employ. Psycho-metrics emphasizes the use of multiple measures to assess underlying traits that are not directly observable. Econometrics emphasizes approaches to time-series and multiple-time series of observational data. Both of these branches take the individual case as the unit of analysis. The distinctive feature of socio-metrics arises from its emphasis on the relation between cases, rather than the attributes of individual cases, as the unit of analysis.

SNA is a particular application of the analysis of relational structures to patterns of ties among social actors. SNA has developed quite a large toolkit for the description of the distributions of relations among actors, such as degree-distribution, centrality, clustering, path-length, etc. (see, for example, Hanneman and Riddle, 2005; Kadushin, 2012; Wasserman and Faust, 1994). Additionally, a great deal of work has been done over the past 50 years on modeling and hypothesis testing of social network data.

Much of the work on statistical analysis of social networks focuses on relationships among two or more networks, or the change in a single network over repeated observations (the dynamics of networks). Other work integrates the analysis of network data with the analysis of data on the attributes of the individuals who make up (or are "embedded in") the network. Sometimes the network plays the role of independent variable in analysis of how the attributes of other actors influence the attributes or behavior of a focal actor. Sometimes networks are taken as the dependent variable; i.e. the selection of a particular pattern of relations among the actors is seen as arising from the attributes of the embedded actors. Recent work has explored the "co-evolution" of the distributions of individual actor attributes and distributions of dyadic (or relational, or network) ties.

The text will first introduce the most common data structures used in the statistical analysis of social network data. It will then explore the analysis of attribute data, when the cases being observed are embedded in a network. We start with the analysis of symmetric bivariate association, move to asymmetric association in which either the network or the attribute may be the dependent variable. From there, we move to a more extended treatment of approaches examining the relations between actors as the outcome of interest.

1.5 References

Brieger, Ronald L. 1974. "The Duality of Persons and Groups." *Social Forces* 53(2): 181-190.

Hanneman, Robert A. and Mark Riddle. 2005*. Introduction to Social Network Methods*. http://faculty.ucr.edu/~hanneman/nettext.

Huisman, Mark and Marijtje A J Van Duijn. 2011. “A Reader’s Guide to SNA Software,” Pp. 578-600 in *The SAGE Handbook of Social Network Analysis*, edited by J. Scott and P. J. Carrington. London: Sage.

Kadushin, Charles. 2012. *Understanding Social Networks: Theories, Concepts, and Findings*. Oxford: Oxford University Press.

Moreno, Jacob L. 1951. *Sociometry, Experimental Method, and the Science of Society*. Ambler, PA: Beacon House.

Wassermann, Stanley, and Katherine Faust. 1994. *Social network analysis: Methods and applications*. Cambridge, UK: Cambridge University Press.

Chapter 2. Working with Relational Data

In this chapter, we will look at some of the main ways in which the data for SNA are structured for use in statistical analysis. Our examples will use UCINET, which is good for many basic statistical analyses of network data (on fairly small data sets). We will also briefly discuss how network data can be structured for use in multi-level modeling and exponential random graph modeling.

Many of the ideas and the terminology of the statistical analysis of social network data can be a bit difficult when discussed in abstract conceptual terms. However, they become clearer when we apply them to actual data. We will start the chapter by taking a look at the dataset that will be used for most of our examples throughout the text.

2.1 About the Example Data

For most of our examples here, we will be using some data about the students in an upper division, undergraduate course in social networks (taught in the fall term of 2011). The data are a population census (not a sample, or set of ego-networks). The students were asked to report about acquaintanceship (i.e. "who in the class do you know well enough to ask a small favor, like borrowing class notes?"), on four occasions over the 11-week class. That is, the design is a panel, rather than a single-cross section, or continuous time set of observations. Information was also collected about some "fixed" attributes (ethnicity, gender), and about some "time varying" attributes (attendance, grades on examinations) that occurred over the term. The acquaintanceship networks for the four waves are shown in figures 2.1 through 2.4, colored by ethnicity (blue = White; red = Hispanic; green = Asian; yellow = African-American), with men shown as circles and women as squares.

Figure 2.1. Acquaintanceship Network for Classroom Data, First Day of Class

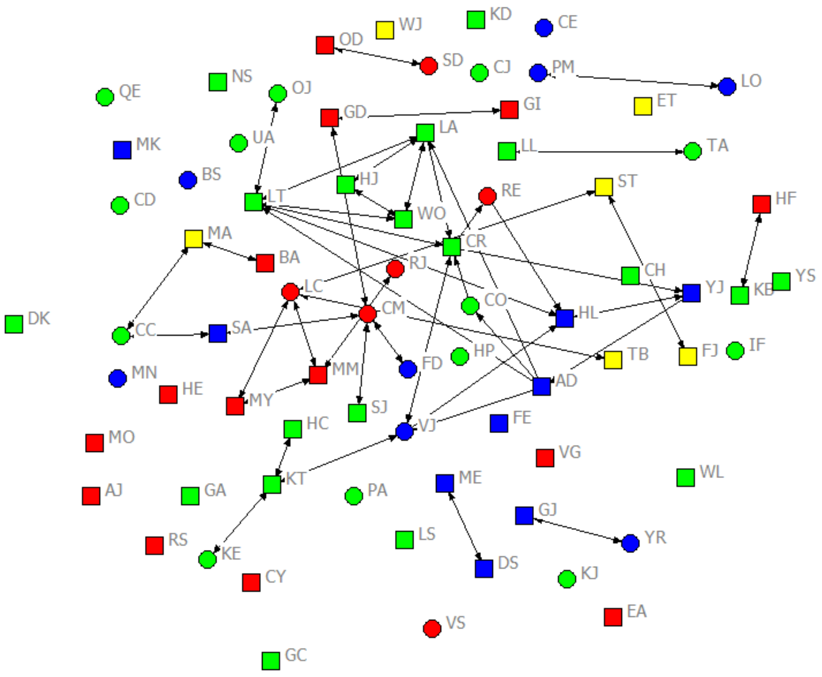


Figure 2.2. Acquaintanceship Network for Classroom Data, First Mid-term Exam

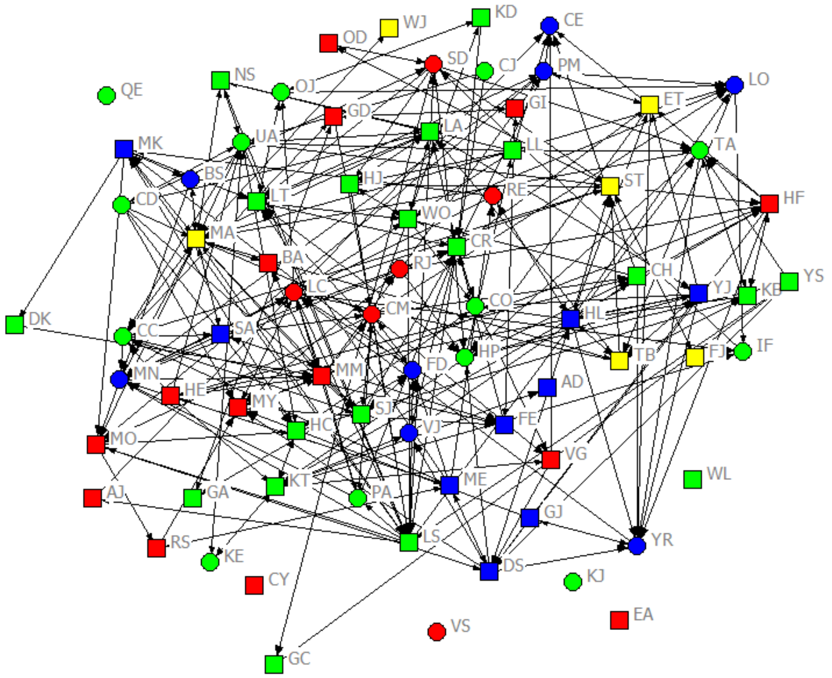


Figure 2.3. Acquaintanceship Network for Classroom Data, Second Mid-term Exam

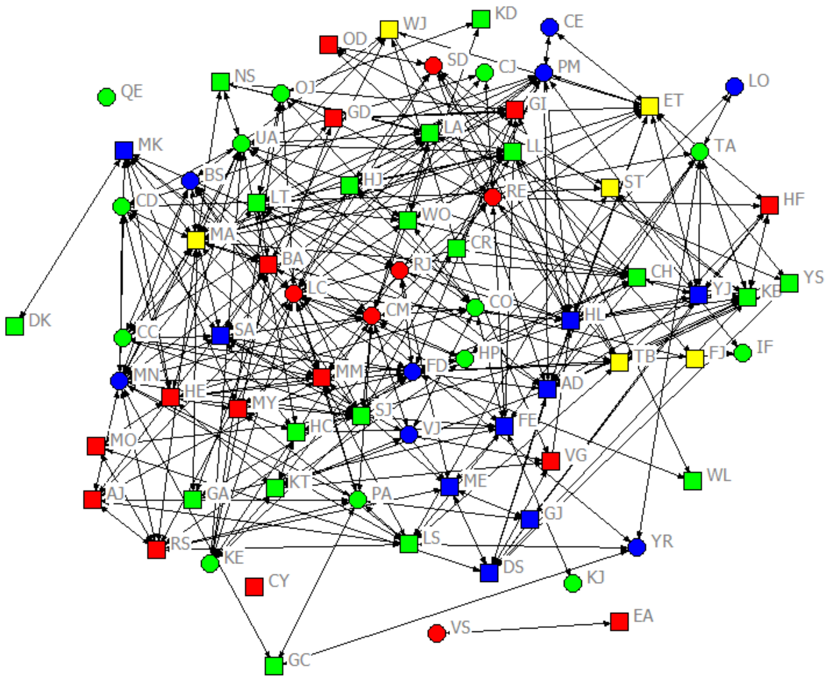
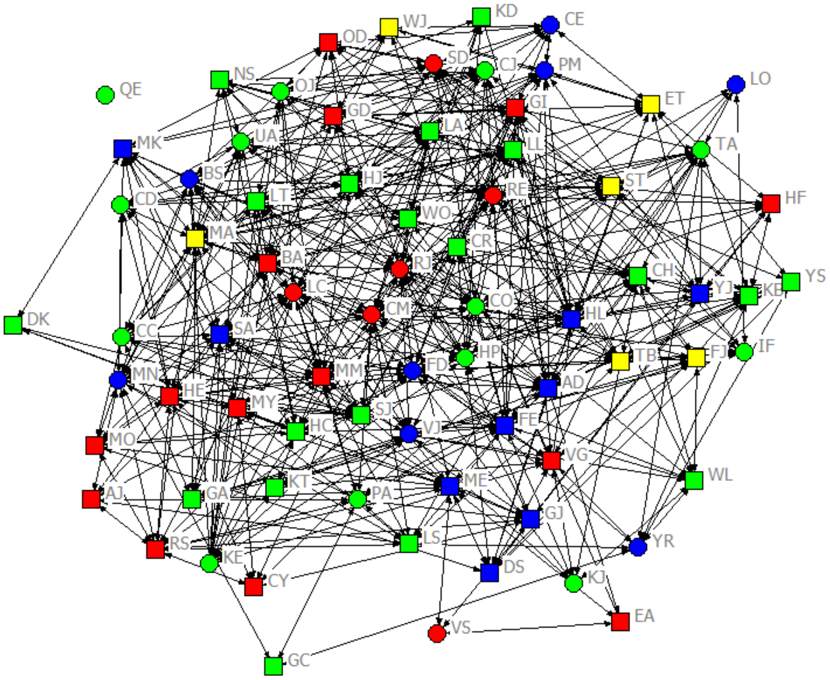


Figure 2.4. Acquaintanceship Network for Classroom Data, Final Exam



More complete details about the classroom data are available on this page. Readers will also be able to download the data there as well in both UCINET and Excel formats. The text was written so that readers can follow the examples by recreating them on their own.

There were seventy-five students in the class. Over the academic term, the density of ties increased quite a lot. On the first day of class, there were 78 unique ties (Figure 2.1). By the first midterm, there were 315 (Figure 2.2). By the second midterm, there were 530 (Figure 2.3). And at the time of the final exam, there were 895 ties (Figure 2.4). Notably, not all students who said that they were acquainted had reciprocated ties. For example, in the center of Figure 2.1, we can see that CO indicated that they knew CR, but CR did not acknowledge CO as an acquaintance. In the same figure, AD claimed to know CO, but CO did not acknowledge knowing AD. Network ties that are not necessarily reciprocated are referred to as "directed" ties. There were more women than men in the class, but reasonable numbers of each gender. The class had considerable ethnic diversity, with more students identifying as Asian than any other group.

SNA analysts use a wide variety of research designs, sampling, and measurement. Design, sampling, and measurement choice all greatly affect how the resulting data need to be analyzed. We can't hope to be comprehensive in this text. This data set lets us illustrate the kinds of questions and approaches that are most common in current social science work. Understanding the basics with a fairly simple and familiar example is a starting point for analysts with more complicated problems.

Most of the examples in this textbook are done in UCINET, which is why we made the data available in UCINET data files. UCINET data are stored in pairs of files, one with the extension .##h and the other .##d. You can call either one to open the data. The .##d file contains the actual data, and the .##h file contains information on how to call and read the .##d file. You will need to download and store both in the same folder to open a dataset. One set of files used for examples throughout this text is titled *attributes2011*, which holds the student attribute data. The other four sets of files (*wave1_2011* through *wave4_2011*) contain the acquaintanceship data indicating which students nominated which other students as acquaintances.

2.2 Attribute (Nodal) Data Structures

What is unfamiliar about the statistical analysis of social network data is its focus on relations between cases as the "unit of analysis," rather than the more familiar focus on attributes of individuals (which vary across individuals, and are called "variables" outside SNA).

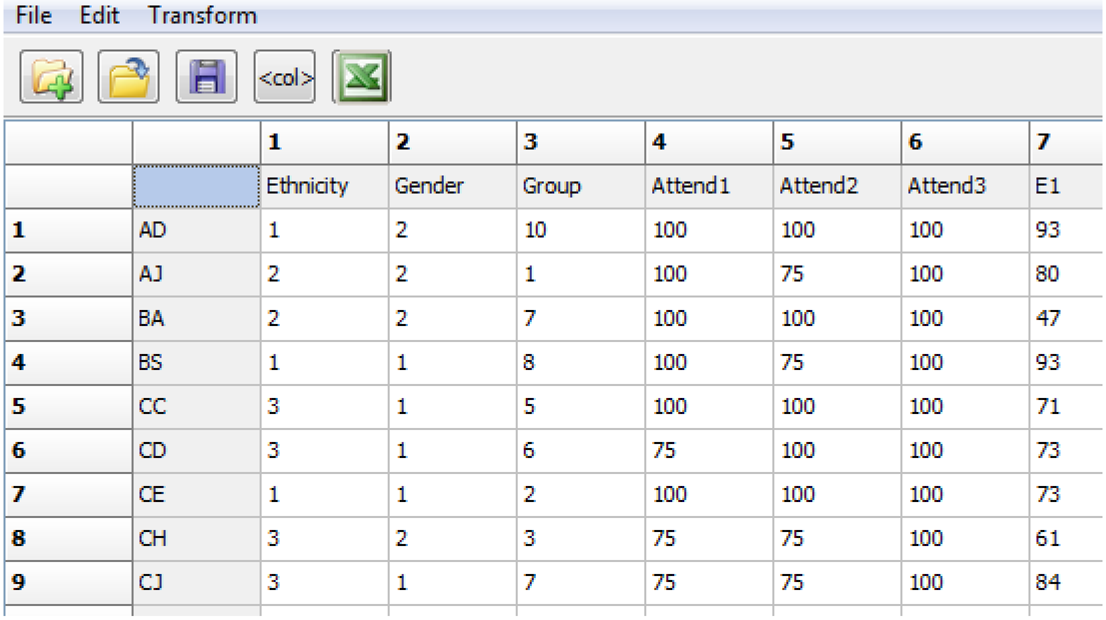
SNA does work with variables, though. It thinks about them as "attributes" of nodes. For example, a conventional, variable oriented analyst might say that "Susan has the score or value of woman on the variable gender." A social network analyst would say "the *node* Susan has the *attribute* woman."

Sometimes attributes play the role of independent variables used to explain or predict network variables. For example, are women students in our class ("woman" being an attribute of each individual) more likely to be acquainted with other women than they are with men?

Sometimes attributes play the role of dependent variables, predicted by network variables. For example, is a student's performance on a test (an individual attribute) predictable from the scores on the test of those the student is acquainted with (an observation about dyads)?

Recording data about the attributes of individuals for use in UCINET and other network software is quite familiar. Data are arrayed in the conventional "rectangular" way (cases by variables) or "vector" (cases by a single variable). Figure 2.5 shows a portion of the attributes dataset for the student data. The display is a screen-shot of the dataset viewed through the UCINET Matrix Editor (note: data arrays like this can be created in any software and saved as text files to be imported into UCINET later).

Figure 2.5. Partial View of the Student Attributes Rectangular Data Array in UCINET



		1	2	3	4	5	6	7
		Ethnicity	Gender	Group	Attend1	Attend2	Attend3	E1
1	AD	1	2	10	100	100	100	93
2	AJ	2	2	1	100	75	100	80
3	BA	2	2	7	100	100	100	47
4	BS	1	1	8	100	75	100	93
5	CC	3	1	5	100	100	100	71
6	CD	3	1	6	75	100	100	73
7	CE	1	1	2	100	100	100	73
8	CH	3	2	3	75	75	100	61
9	CJ	3	1	7	75	75	100	84

In UCINET, it is best to make sure that each case's ID is short and a single string, preferably without special characters or spaces. Cases need to be sorted in the same order in the attribute file or files as they are in the relational (dyadic) datasets discussed shortly. It's good to keep a codebook, as UCINET procedures often ask you for a variable number in the attribute set (e.g. *Ethnicity* is col. 1, *Group* is col. 3).

All the standard approaches to coding categorical, ordinal, and interval level variables apply. It is usually best to do most data transformation and elaborate coding outside of UCINET (say in a spreadsheet or statistical package). UCINET does offer some special tools for working with attribute data, however, that we will discuss below. Before turning to these tools, let's briefly discuss some categories of attribute variables that are used in SNA.

Most specialized social network analysis software systems record information about nodal attributes in the same general way as UCINET. However, each attribute is generally stored in a separate text file (with the cases sorted in the same order across all the files, and varying header information). When relational data is analyzed using mixed-effects generalized linear models, each data line refers to a relation (dyad), and the attributes of the nodes associated with the dyad are coded with each data line.

2.2.1 Fixed and Time-varying Attributes

Some attributes that we might be interested in analyzing are ascribed characteristics of individuals that are fixed for each node. For example, in our data set, a student's gender and ethnicity are treated as static across the four waves. In the statistical analysis of network data, fixed attributes often are used to partition the network (divide it into groups of cases with the same attribute). In mixed-effects modeling, attributes are "level 2" variables (because dyads are the primary unit of analysis – level 1), and dyads are nested in the crossing of two individuals.

Other attributes of individuals may vary over time. In our example, exam score (E1, E2, E3), attendance, term paper score, and term paper team participation all occur at particular points in time. Some might be treated as repeated events or latent growth curves (exam scores or attendance, for example). Time varying covariates may be treated as causes of how a person selects their networks (do students who perform badly on exam 1 create more ties with students who performed better?). Alternatively, time varying covariates might be treated as outcomes of network influence (e.g., does a student's test score on the second exam depend on the test scores of the others in their network at the time of the first exam?).

All in all, these kinds of variables or attributes are entirely familiar. There is another kind of attribute data used in SNA that might not be quite as obvious.

2.2.2 Network Position as an Individual-level Attribute

One of the key ideas of SNA is that how an individual is "embedded" in the network may affect either their attributes and behavior, or their selection of social relations (making and breaking ties). For example, one important hypothesis is that of "preferential attachment." This principle says that individuals who have more network ties are more likely to be sought as partners than those who have fewer ties. In this case, the number of ties that an actor has is viewed as a variable or attribute of the actor that affects the probability that they will form more ties.

It is very common for analysts to calculate measures of an individual node's position in the network and save these as attributes of the node for further analysis. The specific aspects of a node's position that might be relevant, of course, will vary with the goals of the analysis. A node's degree, their centrality, the clustering of their ego-network, and the homophily of their ties (the proportion of their ties that are ties to others with the same attribute as themselves) are common things about an individual's network position that are often treated as nodal attributes in further analysis. While such variables describe a node's position in a network, the variables are actually an attribute of the node, and can be treated as such.

When working with UCINET, most procedures (for example, calculating the between-ness centrality) automatically output a dataset containing the case labels and vector(s) of results. These output files can be used directly as input in other routines, or they can be appended to an existing attribute file using the "Join" procedure (discussed below and in the next chapter).

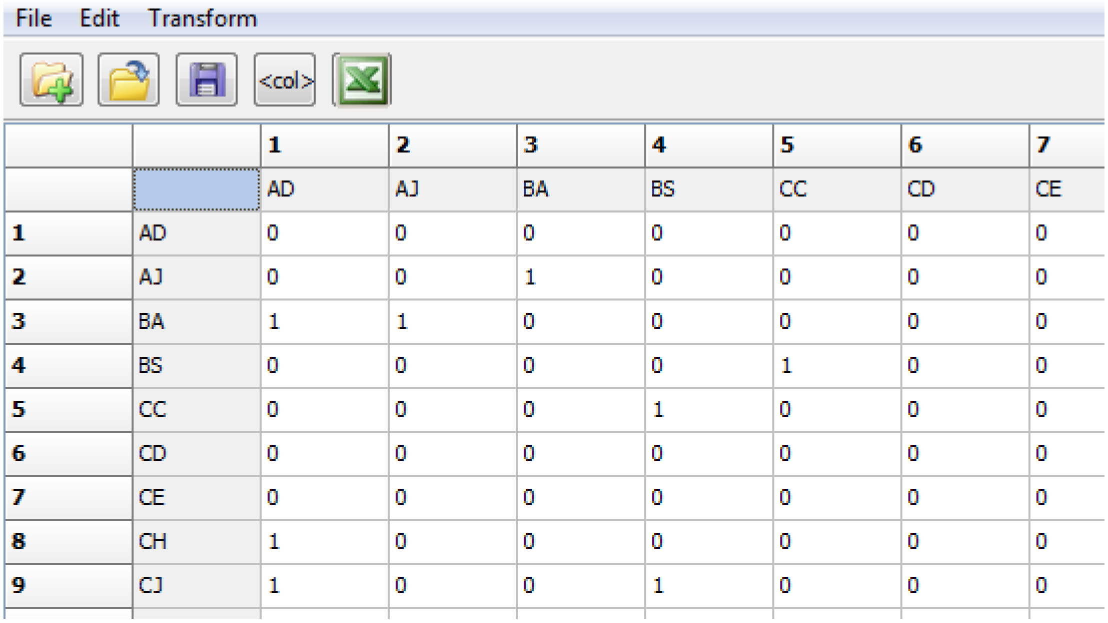
One important part of network data sets then, are arrays that describe attributes of nodes. Attribute data are familiar; they record the scores on variables (i.e. the attributes) of cases (nodes). They can be stored as single vectors (with scores for each node on a single attribute), or as lists-of-lists in a rectangular matrix. Most specialized network analysis programs (e.g. R-Statnet or Sienna) will want each attribute stored in a separate data-file. Statistical packages (e.g. Stata) will want rectangular arrays.

Now, let's turn our attention to the less familiar notion of relational, dyadic, or network data.

2.3 Relational (Dyadic) Data Structures

A network is made up of nodes and the relations between them. The basic building-block of a network, then, is the relation between two actors. When all the dyads are collected together, larger and more complex structures emerge. "Relational" or "dyadic" or "network" data describe the relations among all pairs of actors. Figure 2.6 shows a portion of the fourth wave of "acquaintanceship" among our social networks students, as an example.

Figure 2.6. Partial View of the Acquaintanceship Data in UCINET, Wave 4



		1	2	3	4	5	6	7
		AD	AJ	BA	BS	CC	CD	CE
1	AD	0	0	0	0	0	0	0
2	AJ	0	0	1	0	0	0	0
3	BA	1	1	0	0	0	0	0
4	BS	0	0	0	0	1	0	0
5	CC	0	0	0	1	0	0	0
6	CD	0	0	0	0	0	0	0
7	CE	0	0	0	0	0	0	0
8	CH	1	0	0	0	0	0	0
9	CJ	1	0	0	1	0	0	0

In the example, we see that node “BA” said that they were acquainted with node “AD.” But, note that “AD” does NOT indicate that they are acquainted with “BA” (read across the rows). Relational data are stored in square node-by-node matrices. Data may be “asymmetric” or “directed” (like the above example) where A→B does NOT necessarily imply B→A. In asymmetric data, the source of the relation is the row, and the destination of the relation is the column. Data may also by “symmetric” or “un-directed” or “bonded” where A→B *does* imply B→A. The diagonal cells (AD to AD, etc.) are usually ignored in SNA. It is conventional to code them as zero.

Social networks that have large numbers of actors can produce very large square matrices which can be difficult to edit, store, and process. Another very common way of recording relational data is the “edge list.” An edge-list is a list of the dyadic relations that are present in the social network (leaving out those that are absent). An edge-list contains the identity of the origin node and destination node for a directed tie (or the two nodes involved, in any order, in a symmetric tie), and the value of the tie. Many network analysis software packages are able to work with edge-list data (they convert it to full matrices for analysis). Statistical packages, when working with relational data, usually define each row of data as an “edge” (but they will also require a listing of the edges or relations that are absent).

The entries in a relational dataset may be binary, multinomial, ordered, or interval-ratio. Relations between the nodes in each dyad commonly represent the presence/absence of a tie, or what “type” of tie exists, or the strength or probability of a relation. Appropriate statistical treatment of relational data, of course, depends on the way that the relational variable has been measured. Most of our examples will be very simple, focusing on the simple presence or absence of a tie.

Individual nodes in SNA can have any number of attributes. These are stored in the dataset as one or more rectangular (node by attribute) matrices, or one or more vectors. Similarly, a SNA dataset can have any number of relational variables. Let’s consider some of the major types of relational variables used in many SNA applications.

2.3.1 Building Relational Variables from Attributes

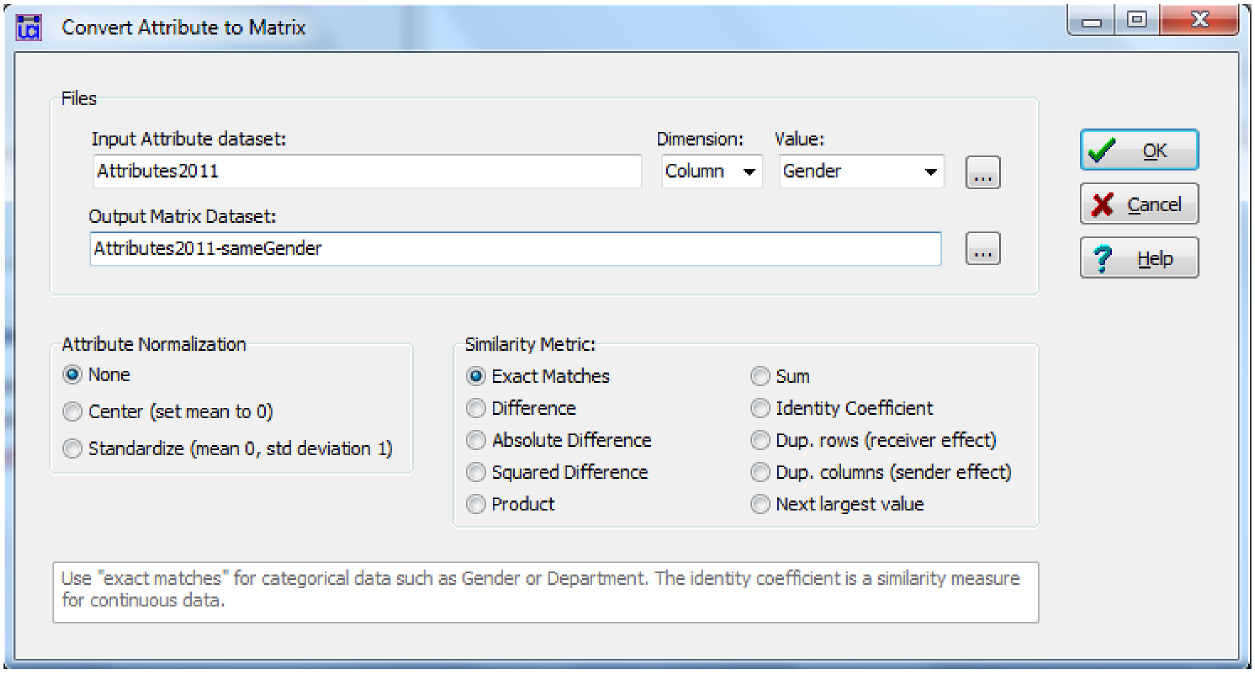
In SNA, our focus is often on the existence, strength, or quality of the relationship between two nodes. Sometimes it is helpful to characterize the relationship between two actors as a function of the comparison between the attributes of each.

In SNA, the relationship between two nodes is "nested" within the pair of nodes. As an example, suppose that Fred is a man and Susan is a woman. We might wish to characterize the relation between Fred and Susan as being between two persons that differ on gender. That is, the dyadic relation between Fred and Susan is not gender homophilous.

The attributes of individual nodes are often important in SNA, but the more common SNA questions concern the comparison of the attributes of two nodes, which describes the dyad. One can imagine an analysis that uses attributes in both ways. For example, are women (individual level variable) more likely than men to have ties to persons of the same gender (a relational characteristic)?

UCINET has an interesting tool for comparing the attributes of two individual actors and building a dyadic variable (a relational variable that exhibits properties of dyads typically displayed in matrix form like the data in Figure 2.6) to describe the relation between them. Consider the dialog box created by *Data>Attribute to matrix*, shown as figure 2.7.

Figure 2.7. UCINET Dialog for Converting the Attributes of Pairs of Nodes to a Dyadic Variable



Files

Input Attribute dataset: Attributes2011

Dimension: Column

Value: Gender

Output Matrix Dataset: Attributes2011-sameGender

Attribute Normalization

● None
○ Center (set mean to 0)
○ Standardize (mean 0, std deviation 1)

Similarity Metric:

● Exact Matches
○ Difference
○ Absolute Difference
○ Squared Difference
○ Product
○ Sum
○ Identity Coefficient
○ Dup. rows (receiver effect)
○ Dup. columns (sender effect)
○ Next largest value

Use "exact matches" for categorical data such as Gender or Department. The identity coefficient is a similarity measure for continuous data.

In this dialog, we've identified our rectangular attribute data set as input. We've selected the column (rather than the row, because our data array is actors-by-attributes, not attributes-by-actors), and the individual attribute "Gender." The output of this procedure will be stored in a new dyadic (node-by-node square matrix) data set. In this case, we've selected "Exact Matches" in the "Similarity Metric" panel. This means: if the score of node A on the variable gender is the same as the score of node B on the variable gender, then code a "1" in the new matrix element AB (and BA). That is, the result of this operation is to build a matrix with "1" if two nodes are the same gender, and "0" if they are not. That is, a matrix of "gender homophily" is created.

There are a number of other useful functions here. "Difference" generates a "1" if two nodes are dissimilar. Absolute and squared differences measure the distance between two nodes on a quantitative attribute. Product and Sum generate a characterization of the dyad that aggregates their attributes. For example, we might think that the importance of the tie between two actors A and B is the synergy or multiplication of the degree of each of the two individuals.

The dialog also provides for normalizing the results (usually used with quantitative attributes). One may also automatically center the resulting dyadic score. This is often used in multi-level analysis to center the scores on the "level-1" (dyadic) variable around their mean.

2.3.2 Repeated Measures

SNA work has increasingly focused on studying the processes that generate and modify network structures over time. While some research designs and data sources allow us to identify the exact time at which each dyadic relation is created or modified, most data sets consist of panel data – observations on the state of the dyadic ties as multiple, discrete, points in time. Our example data set, for instance, observes the ties among pairs of 75 students at four discrete times during an academic term.

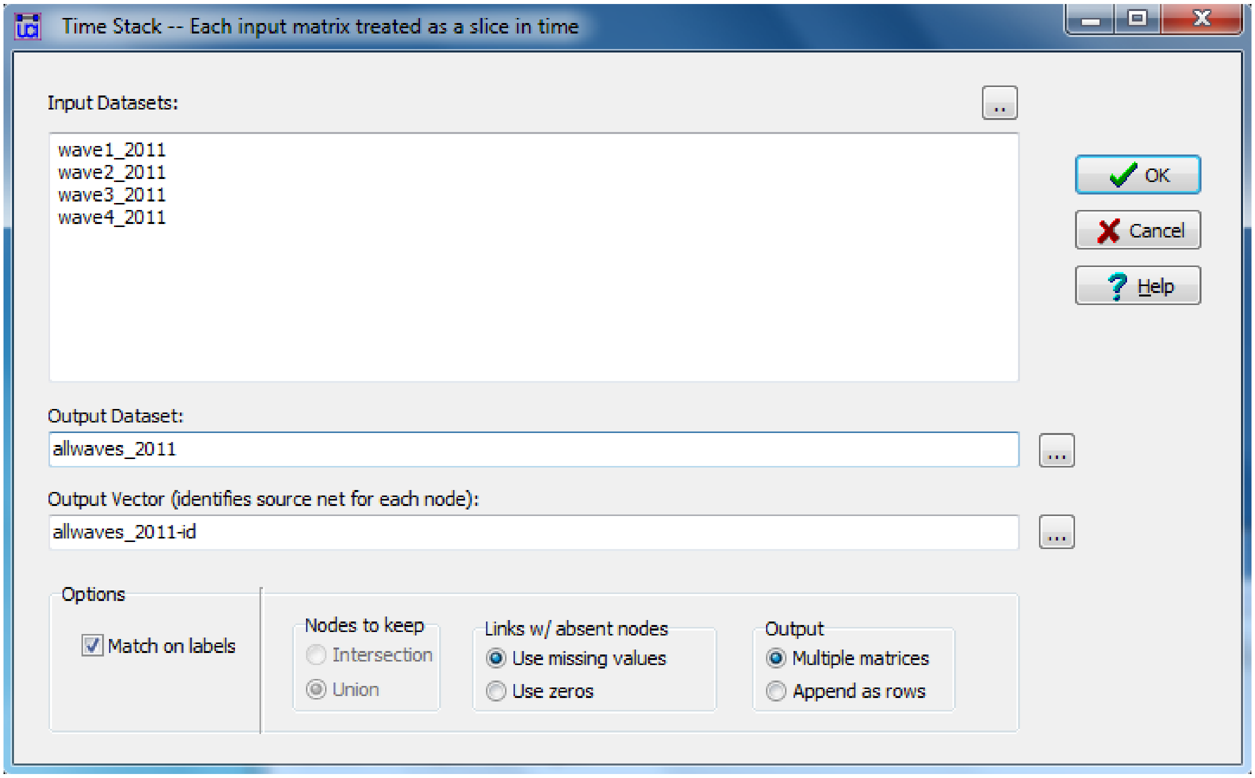
Data on the same relation among the same actors is stored as a series of square node-by-node matrices in UCINET (with identical node labels and the nodes in the same order). For convenience in processing and displaying repeated dyadic measures, the matrices can be "stacked" into a single file.

Data>Join>Join Matrices combines or stacks multiple matrices with the same rows and columns into a single file. Subsequent runs in UCINET procedures will process all the matrices in the stacked file, and produce analyses for each matrix.

Another UCINET tool can be used when the multiple observations are missing some of the actors. This circumstance might arise where the data are taken from observations or documents at different time points, each of which reports on the actors present at the particular time of observation. Social network data are often collected by querying the actors who happen to be present at a particular time point, and each matrix may not include all of the actors who were ever present. If we want to include all actors in the analysis of each panel, we can use the time stack tool to build conformable data sets for each panel.

Transform>Time Stack has a dialog box that looks like Figure 2.8.

Figure 2.8. UCINET Dialog for Merging Dyadic Panel Data



Time Stack -- Each input matrix treated as a slice in time

Input Datasets:

wave1_2011
wave2_2011
wave3_2011
wave4_2011

Output Dataset:
allwaves_2011

Output Vector (identifies source net for each node):
allwaves_2011-id

Options

☑ Match on labels

Nodes to keep
○ Intersection
● Union

Links w/ absent nodes
● Use missing values
○ Use zeros

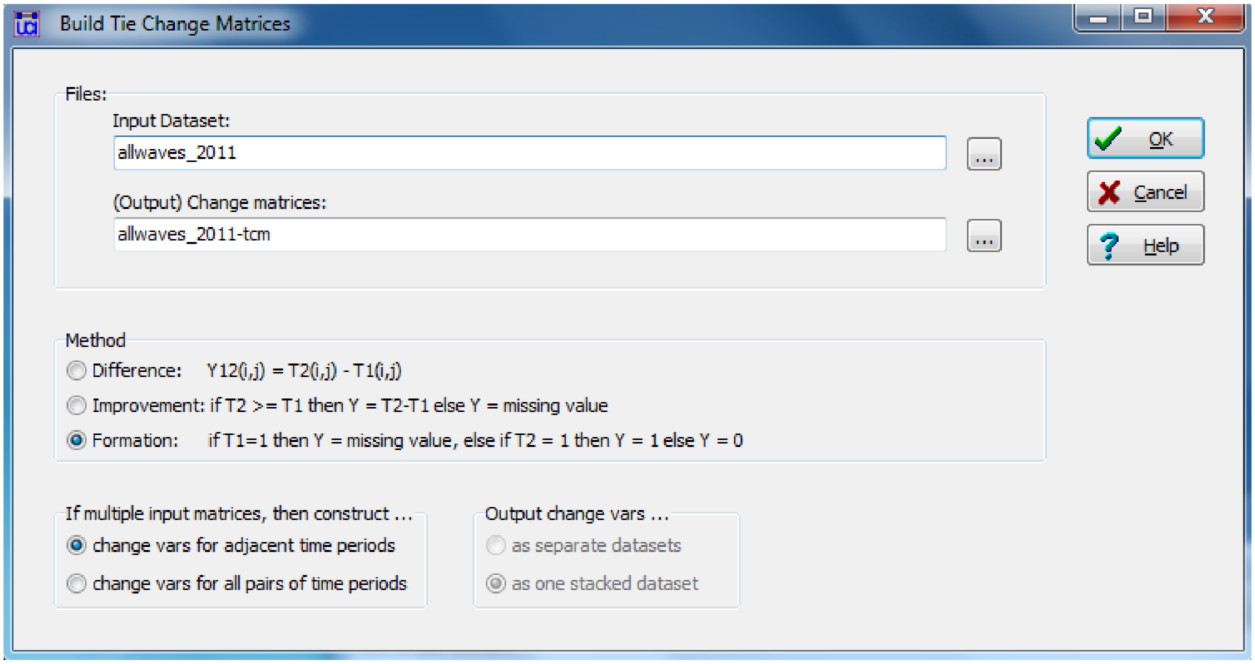
Output
● Multiple matrices
○ Append as rows

We've identified the student-by-student acquaintanceship at each of the four time points as our source data, and chose to save the data in a new file called "allwaves_2011" for future use. The default "Match on labels" has been selected for UCINET to determine which cases match across files. We've asked that missing value codes be entered for cases that are not present in a particular file (alternatively, one might want to assume that the case was present, but had no ties, by selecting "use zeros"). The output file or files will contain rows for all nodes that are in any of the input files.

When repeated measures have been joined into a stacked matrix in this way, another tool can be used to create a new dyadic variable that identifies change in ties. Tie change matrices can be interesting descriptively (e.g. did tie additions, i.e. new acquaintanceships, accelerate in the period of cramming for the final exam compared to other times in the quarter?). Tie changes can also be used as a dependent variable to test hypotheses about network change (e.g. are women more likely than men to form new ties?).

Transform>Build tie change matrices generates a dialog like that in Figure 2.9.

Figure 2.9. UCINET Dialog for Generating Tie Change Data



Files:

Input Dataset: allwaves_2011

(Output) Change matrices: allwaves_2011-tcm

Method

○ Difference: Y12(i,j) = T2(i,j) - T1(i,j)

○ Improvement: if T2 >= T1 then Y = T2-T1 else Y = missing value

● Formation: if T1=1 then Y = missing value, else if T2 = 1 then Y = 1 else Y = 0

If multiple input matrices, then construct ...

● change vars for adjacent time periods

○ change vars for all pairs of time periods

Output change vars ...

○ as separate datasets

● as one stacked dataset

The input dataset is our time stacked data, and we can select a name for the matrices that result from the operation. The "Method" choices allow generating measures of any "difference" between two matrices (and it's amount, if the relation is quantitative); or, we can select "improvement" to record only positive changes; or we can select "formation" to capture cases where there was no tie at time one, but a tie was formed by time two. One can also select whether the resulting stacked change matrices are calculated for adjacent points in time only, or for all pairs of time points.

2.3.3 "Training" Networks

The relationship between two actors may be a function of a different relationship between them. In our example data, the students in the class were placed (more or less randomly) into groups to work on research term papers. An obvious hypothesis would be that being placed into the same research team will result in becoming acquainted with other team members (actually, this didn't always happen in the class!).

This is an example of one network (who shares the relation of "being in the same work group with") that "trains," or acts as a "context" or "social geography," that may guide the formation of new acquaintanceship ties. Being present in the same physical space, being members of the same organization, and other forms of affiliation can modify the likelihood of ties of other types forming.

Similarly, if one type of tie exists between two nodes (say, one goes to the other for advice on the job), this may modify the likelihood of the formation of another kind of tie (say, being friends outside of work). In a slightly different sense, the presence of one kind of network tie (advice seeking) is "training" the other network (being friends).

Network data that describe the presence of joint affiliation, or context, or other kinds of ties are recorded just like any other network or relational variable – as a square matrix (node-by-node, directed or not, valued or not). As we will see later, the relation between two nodes in one (training) network can easily be used to predict the relation between the two nodes in another.

2.4 Affiliation

One of the most important strengths of the way that SNA conceptualizes social structure is in recognizing how contexts shape (and are sometimes shaped by) patterns of ties between pairs of social actors. The most often cited example of this idea is the study by Davis, Gardner, and Gardner (1941) of sociability among a group of women in a southern town. Davis et al. recorded whether each of the women had attended each of a number of social events. The strength of ties between pairs of women can be indexed by the number of events they both attended (and/or both did not attend).

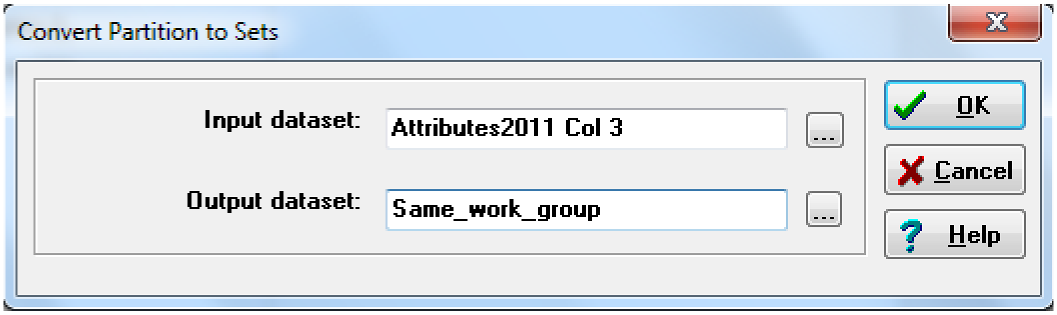
Data structures that show the ties between two types of social entities (in the Davis et al. example, women and social events) are called "two-mode" (or "bi-partite" or "affiliation"). The data are usually recorded as a rectangular matrix of actors (on the row) by events (on the column). For use in our statistical analyses where we are predicting dyadic outcomes, affiliations need to be turned into node-by-node matrices. The idea is simple: the tie between two nodes is some function of the pattern of their affiliations. Usually, if two actors have very similar affiliations, we regard the tie between them as strong; if they have very dissimilar affiliations, we regard their tie as weak.

In many social network analyses, it can be very important to create dyadic variables that measure shared context or joint affiliation between the members of dyads. Such variables often represent the similarity of the locations of the actors in "social space" and can be strong predictors of the likelihood of other kinds of social ties between them.

There are a variety of approaches and tools for turning affiliation data into dyadic variables for use in SNA. Here are just a few:

Sometimes we have stored information about context or affiliation as a vector. In our student data, for example, we recorded which of 10 workgroups a student was assigned to. We might want to turn this into a student-by-student dyadic variable, coded as "1" if two students were in the same workgroup and "0" otherwise. The UCINET tool *Data>Partition to sets* will help to accomplish this task, as in figure 2.10.

Figure 2.10. UCINET Dialog for Partition to Sets

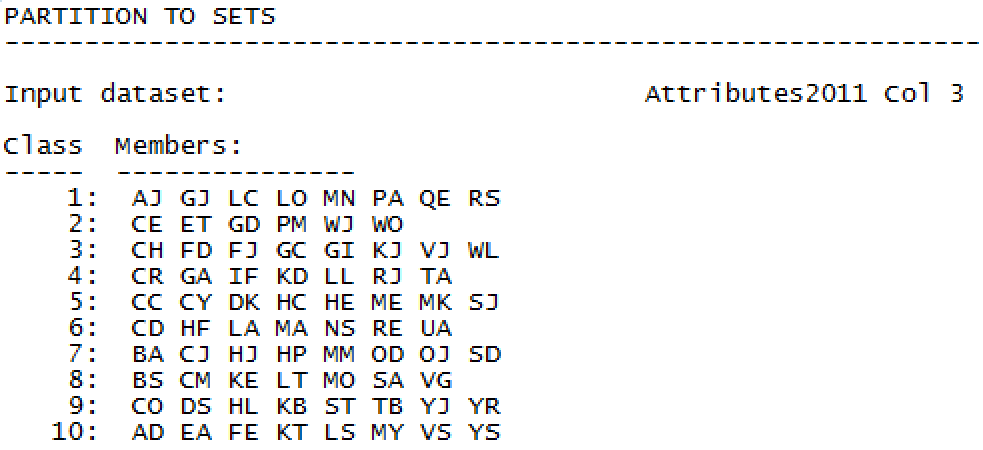


Input dataset: Attributes2011 Col 3

Output dataset: Same_work_group

In this dialog, we've opened our "attributes" file (nodes by attributes), and told UCINET to use column 3, which (from our codebook) we know to be the "work group number" variable. We've given a new name to the output file. The output dialog is shown in Figure 2.11.

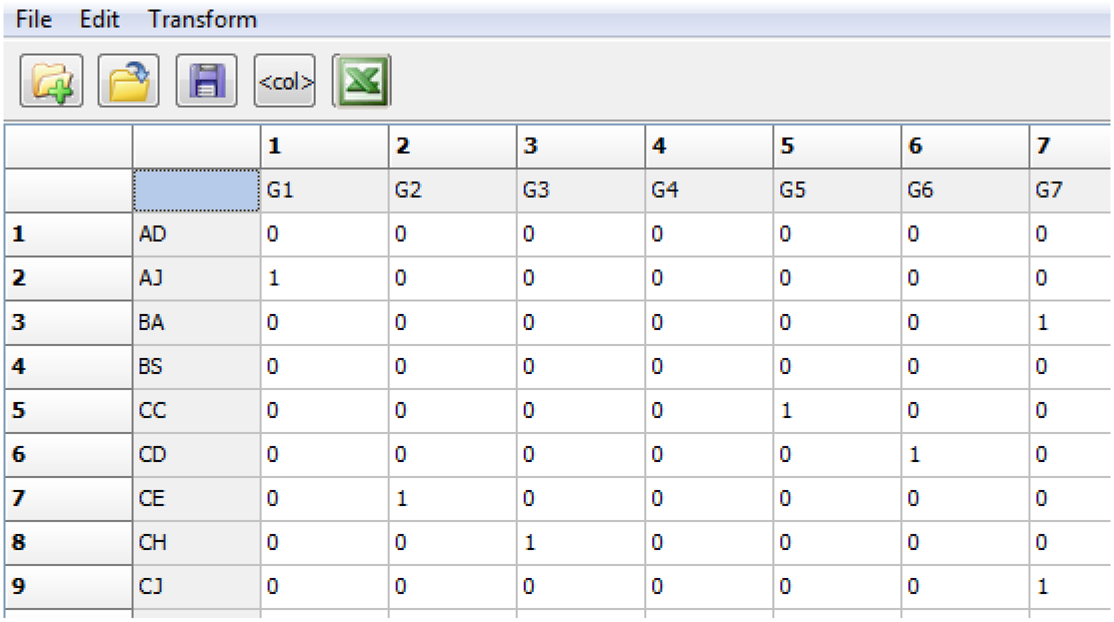
Figure 2.11. UCINET Output of Partition to Sets



```
PARTITION TO SETS
------------------------------------------------------------

Input dataset:                        Attributes2011 Col 3

Class  Members:
-----  ---------------
    1:  AJ GJ LC LO MN PA QE RS
    2:  CE ET GD PM WJ WO
    3:  CH FD FJ GC GI KJ VJ WL
    4:  CR GA IF KD LL RJ TA
    5:  CC CY DK HC HE ME MK SJ
    6:  CD HF LA MA NS RE UA
    7:  BA CJ HJ HP MM OD OJ SD
    8:  BS CM KE LT MO SA VG
    9:  CO DS HL KB ST TB YJ YR
   10:  AD EA FE KT LS MY VS YS
```

The output data file that we called "Same_work_group", a portion of which is shown in Figure 2.12, is a node by attribute "affiliation" matrix. That is, it shows, for each student, whether they were a member of group 1, group 2, …, group 10, with dummy codes.

Figure 2.12. UCINET Output Data from Partition to Sets

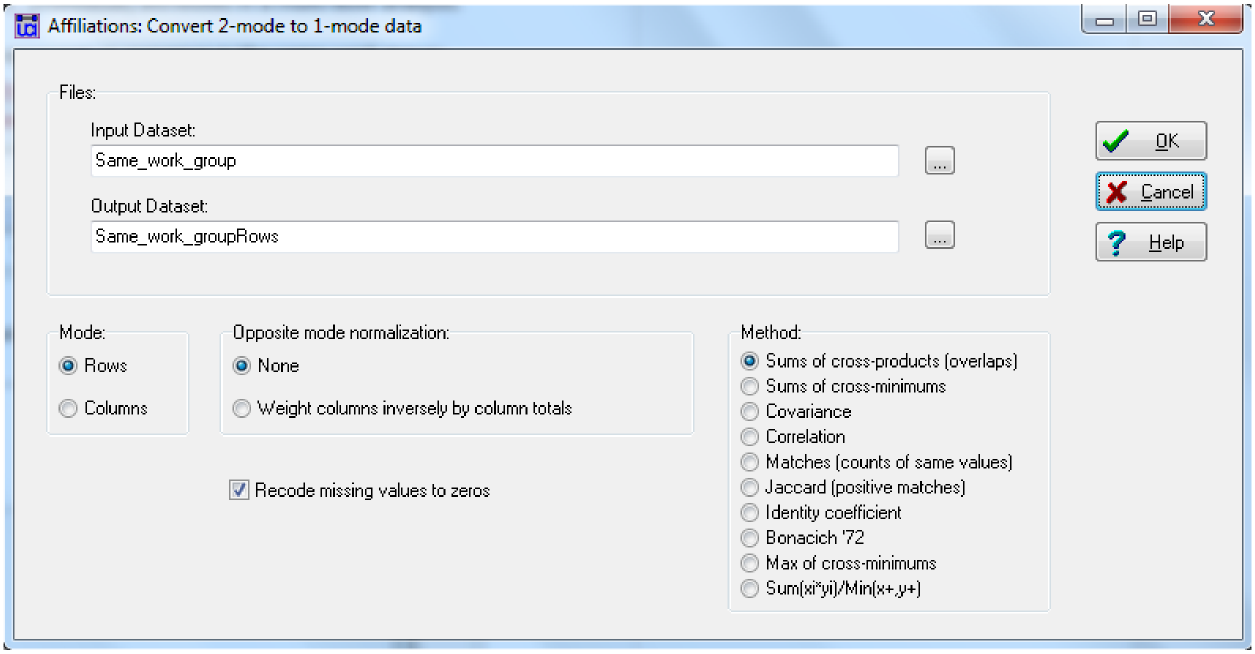


		1	2	3	4	5	6	7
		G1	G2	G3	G4	G5	G6	G7
1	AD	0	0	0	0	0	0	0
2	AJ	1	0	0	0	0	0	0
3	BA	0	0	0	0	0	0	1
4	BS	0	0	0	0	0	0	0
5	CC	0	0	0	0	1	0	0
6	CD	0	0	0	0	0	1	0
7	CE	0	1	0	0	0	0	0
8	CH	0	0	1	0	0	0	0
9	CJ	0	0	0	0	0	0	1

These 10 dummy variables could be treated as "level2" (individual) attributes in a multi-level analysis.

We may also want to create a dyadic variable to represent co-presence in the same work group; this is a dyadic or "level 1" variable. To convert the "Same_work_group" data file to a student-by-student dyadic variable, we can use the *Data>Affiliations (2-mode to 1-mode)* tool. The dialog is shown in figure 2.13.

Figure 2.13. UCINET Dialog for *Affiliations: Convert 2-mode to 1-mode Data*



Affiliations: Convert 2-mode to 1-mode data

Files:

Input Dataset: Same_work_group

Output Dataset: Same_work_groupRows

Mode:
● Rows
○ Columns

Opposite mode normalization:
● None
○ Weight columns inversely by column totals

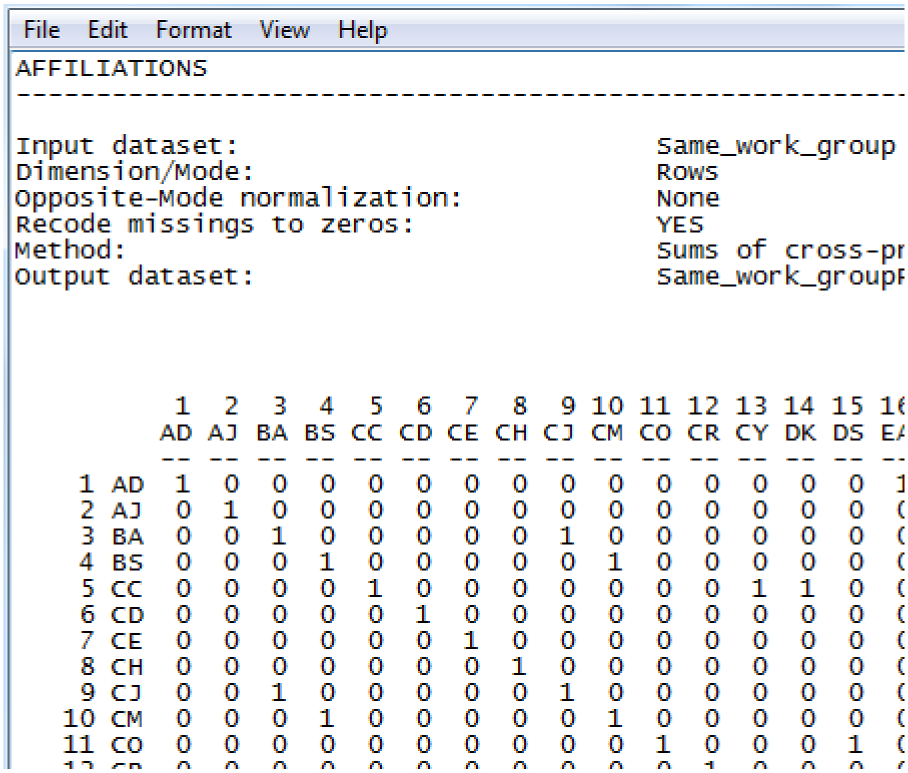
☑ Recode missing values to zeros

Method:
● Sums of cross-products (overlaps)
○ Sums of cross-minimums
○ Covariance
○ Correlation
○ Matches (counts of same values)
○ Jaccard (positive matches)
○ Identity coefficient
○ Bonacich '72
○ Max of cross-minimums
○ Sum(xi*yi)/Min(x+,y+)

We've input our affiliation matrix of students by workgroup dummy variables, and accepted the default output name ("Same_work_groupRows"). This UCINET tool has many useful options. We can choose to work on either rows (to create student-by-student output) or columns (to create a workgroup-by-workgroup matrix). We can normalize output (useful when the input is quantitative rather than binary), and insert zeros for missing data. There are also many useful choices about how to define "affiliation." The "sums of cross-products" multiplies each element of the work-group membership of one person by that of the other person, and sums the result. With binary data, and only one group membership for each actor, this creates a matrix where a pair is coded "1" if they are in the same work group, and zero otherwise. The other options provide alternative ways of operationalizing the idea that two actors have more “similar” affiliations or stronger ties.

The output file that results from this dialog is shown as figure 2.14.

Figure 2.14. UCINET Output of Affiliations Tool



```
AFFILIATIONS
------------------------------------------------------------

Input dataset:                         Same_work_group
Dimension/Mode:                        Rows
Opposite-Mode normalization:           None
Recode missings to zeros:              YES
Method:                                Sums of cross-pr
Output dataset:                        Same_work_groupF



         1  2  3  4  5  6  7  8  9 10 11 12 13 14 15 16
        AD AJ BA BS CC CD CE CH CJ CM CO CR CY DK DS EA
        -- -- -- -- -- -- -- -- -- -- -- -- -- -- -- --
  1 AD   1  0  0  0  0  0  0  0  0  0  0  0  0  0  0  1
  2 AJ   0  1  0  0  0  0  0  0  0  0  0  0  0  0  0  0
  3 BA   0  0  1  0  0  0  0  0  1  0  0  0  0  0  0  0
  4 BS   0  0  0  1  0  0  0  0  0  1  0  0  0  0  0  0
  5 CC   0  0  0  0  1  0  0  0  0  0  0  0  1  1  0  0
  6 CD   0  0  0  0  0  1  0  0  0  0  0  0  0  0  0  0
  7 CE   0  0  0  0  0  0  1  0  0  0  0  0  0  0  0  0
  8 CH   0  0  0  0  0  0  0  1  0  0  0  0  0  0  0  0
  9 CJ   0  0  1  0  0  0  0  0  1  0  0  0  0  0  0  0
 10 CM   0  0  0  1  0  0  0  0  0  1  0  0  0  0  0  0
 11 CO   0  0  0  0  0  0  0  0  0  0  1  0  0  0  1  0
 12 CR   0  0  0  0  0  0  0  0  0  0  0  1  0  0  0  0
```

The output file is a binary, symmetric matrix. We see that BA and CJ are coded as being in the same group, for example. This dyadic variable can now be used to test group difference between workgroups.

The “affiliations” tool is particularly useful when we want to define the similarity, or closeness, or strength of ties between members of a dyad based on the profile of ties that they have with contexts, attributes, identities, or other social objects with which social actors can affiliate.

More generally, we may wish to create dyadic variables describing the strength, closeness, or similarity between members of dyads based on the similarity of their profiles across any set of attributes. For example, one might have information about the location of each actor in "Blau-space" (e.g. their gender, ethnicity, social class, education, etc.; McPherson & Ranger-Moore, 1991). We might wish to reduce this complexity to a single similarity score to characterize the social closeness (the inverse of social distance) between the actors in each dyad. Such similarity is often important in promoting the formation of social ties.

Any number of techniques can be used for this kind of scaling of actors. We begin by comparing each pair of actors, and calculating a similarity or a distance between them based on their scores on multiple attributes. UCINET provides two suites of tools for generating similarity or distance matrices: *Tools>Similarities* and *Tools>Dissimilarities & Distances*.

Once an actor-by-actor similarity (or dissimilarity) matrix is created from the attribute vectors, it can be used directly as a dyadic variable. Similarities and distances between actors can also be further reduced and refined by using clustering and/or scaling (*Tools>Clustering*, *Tools>Scaling/Decomposition*).

2.5 Multiple Relations

When other relational variables serve as independent variables in an analysis of a relational dependent variable, they present no special challenges. For example, we might be interested in the extent to which the relational variable "respondents are the same gender" predicts the outcome "respondents have a reciprocal acquaintanceship relation." But there are occasions when we may be interested in more than one relational variable as dependent. For example, we might seek to explain both "friendship" and "advice seeking" as functions of independent variables.

In conventional statistical analysis, multiple dependent variable problems lead us into the territories of simultaneous equations or structural equation modeling. Unfortunately, specifically tailored applications of these techniques don't exist for relational outcomes (as of this writing).

When the problem calls for the treatment of multiple relational variables as dependent, there are a number of choices, all of which have some difficulties.

One can analyze each outcome separately, using the other outcomes as independent, along with other predictors. Of course, this is a mis-specification, and also does not deal with possible correlated errors across the multiple dependent relational variables.

Alternatively, one can scale the multiple outcomes with mathematical or logical operations. For example, if two actors are both friends and advice givers, we might code the dyad "2"; if either relation is present, we might code "1"; if neither relation is present, we would code "0." Or, we could treat the outcome as the presence or absence of each of 4 types of relations (both ties present, only friendship present, only advice giving present, neither present). Ordered outcomes (such as the 0, 1, 2, 3 coding) or multinomial coding (such as treating each combination of relations as a qualitative type of tie) can be analyzed in some multi-level statistical software that will do hierarchical modeling with non-Gaussian dependent variables.

2.6 Statistical Packages for Network Data

Several groups of researchers have made continuing contributions to the development of statistical analysis of network data. As of this writing, there are a number of very interesting and useful packages that have been made freely available to network modelers (Huisman and van Duijn, 2011). UCINET and several other packages (e.g. Pajek) have excellent suites of tools for describing and working with univariate and bi-variate network and attribute data. Multivariate analysis in the generalized linear modeling framework has been developed in specialized (and free) packages such as Stochnet, Statnet, PNET, ORA, and Sienna (http://www.gmw.rug.nl/~stocnet/StOCNET.htm; https://statnet.csde.washington.edu/; http://sna.unimelb.edu.au/PNet; http://www.casos.cs.cmu.edu/projects/ora/; https://www.stats.ox.ac.uk/~snijders/siena/). Most recently, software development for Exponential Random Graph and related statistical modeling has been prepared for the R environment (https://cran.r-project.org/web/packages/statnet/index.html; https://cran.r-project.org/web/packages/sna/index.html; https://cran.r-project.org/web/packages/ergm/index.html; https://cran.r-project.org/web/packages/RSiena/index.html).

The data structures that we've discussed above are common to all of the specialized packages for the descriptive and predictive analysis of network data – though the details of how data are prepared vary somewhat from package to package.

Many (but not all) multivariate predictive analyses of network data can also be carried out with standard commercial statistical software such as Stata and SAS. Because these packages were primarily designed for non-network analysis, they require a rather different data structure. We'll discuss, and show an example of preparing and analyzing relational data with Stata in Chapter 8.

2.7 Summary

Social network analysis often involves the description and explanation of attributes of nodes. Working with nodal attribute data is similar to conventional statistical approaches that analyze attributes of cases at the individual level. Often, social network analysis will treat network position or structural measures of embeddedness as individual level data using them as predictors or determinants of nodal attributes.

Unlike conventional analysis, social network analysis also involves the description and explanation of relational data. Unlike traditional rectangular data arrays, relational data is often represented in square matrices and contains information about dyads or pairs of nodes. Relational variables often display the presence/absence or strength of ties between pairs of nodes. Additionally, relational variables can be generated using nodal attribute data (e.g. the degree to which both nodes in a dyad share a given attribute) and are often used to represent shared affiliation.

This chapter introduced an example dataset that was used to display different types of network data. The dataset contains acquaintanceship information about 75 undergraduate students from four points in time across a single course term and will be used throughout the rest of this text to demonstrate a variety of social network analytic techniques that deal with nodal attribute data, relational data, and both simultaneously.

2.8 References

Davis, Allison, Burleigh Gardner, and Mary Gardner. 1941. *Deep South: A Social Anthropological Study of Caste and Class*. Chicago: University of Chicago Press.

Huishman, Mark and Marijtje A J Van Duijn. 2011. “A Reader’s Guide to SNA Software,” Pp. 578-600 in *The SAGE Handbook of Social Network Analysis*, edited by J. Scott and P. J. Carrington. London: Sage.

McPherson, J. Miller and James R. Ranger-Moore. 1991. “Evolution on a Dancing Landscape: Organizations and Networks in Dynamic Blau Space.” *Social Forces*, 70(1): 19-42.

Chapter 3. Association Between Attributes in Network Data

Having looked a bit at how attribute and relational data are structured, we are ready to start doing some statistics. In this chapter we will look at a type of analysis that is very familiar: studying the association between two variables. In SNA, this is commonly referred to as studying the association between two attributes of the actors in a network.

3.1 Association Between Attributes in Network Data

The notion of studying the association between two variables (while possibly controlling for others) is the bread-and-butter of conventional statistical analysis. So, it shouldn't be surprising that social network analysts are often interested in the association between two (or more) attributes observed across the actors in a social network. With our example data that was introduced in the last chapter, we could test the hypothesis that women were more likely to have higher participation scores in their research groups than men, for example.

We would like to know how strong such a tendency is in our observations (that is, measure the strength of the association); and, we would like to test the null hypothesis that the observed association was the result of a random process.

In the last chapter, we noted that SNA also thinks about how actors are embedded in networks as attributes of the actors – operating to provide opportunities and imposing constraints on their attitudes and behaviors. For example, individual students who are in highly central positions in the classroom network might be more likely to have higher grades. In this case, the centrality of the student in the network is being thought of as an attribute of the individual. How attributes of nodes are related to their positions in the network are key questions asked in SNA (e.g. are men more likely to have higher network centrality than women?).

Studying the association and partial association among attributes of actors in a network is done with exactly the same tools as are used in studying association among variables. Cross-tabulations, tests for differences of means of two or more groups, correlation and regression can all be applied to describe the strength and form of association between the attributes of actors in a network. The descriptive statistics used for association with the attributes of nodes are exactly the same as the descriptive statistics used to describe the association between variables across cases.

But, when we turn to the question of inference and hypothesis testing, we come to a new issue. The formulae that are used to calculate standard errors and test statistics for variation of variables across cases most commonly assume that the cases are independent replications – and often, randomly drawn from an infinite population. SNA data, most commonly, are not samples but populations. And, SNA data are, by definition, not independent replications. It is precisely the non-independence of the cases that is of central interest to SNA! When we measure the association between two traits or attributes of actors who are connected to one another by social ties, it is quite likely that the social ties might have been created, at least in part, as a result of the attributes of the actors.

In testing hypotheses, the logic is to compare some statistic or parameter (for example, a measure of association or partial association) to how much we would expect that statistic to vary from one set of observations to another just by chance (the standard error). The standard error, or sampling variability, or reliability is a statistic often calculated using standard formulae that assume independence. With social network data, standard errors need to be computed differently. Conventional standard errors will be biased – they may be too big (leading us to incorrectly reject a true hypothesis) or too small (leading us to incorrectly accept a false hypothesis).

Estimating standard errors for non-independent observations is a common problem, and has many solutions. Probably the best known and most widely used approaches are the jack-knife, bootstrap, and permutation methods (Efron, 1981). For SNA, the logic of permutation is quite appealing, and is widely used in UCINET.

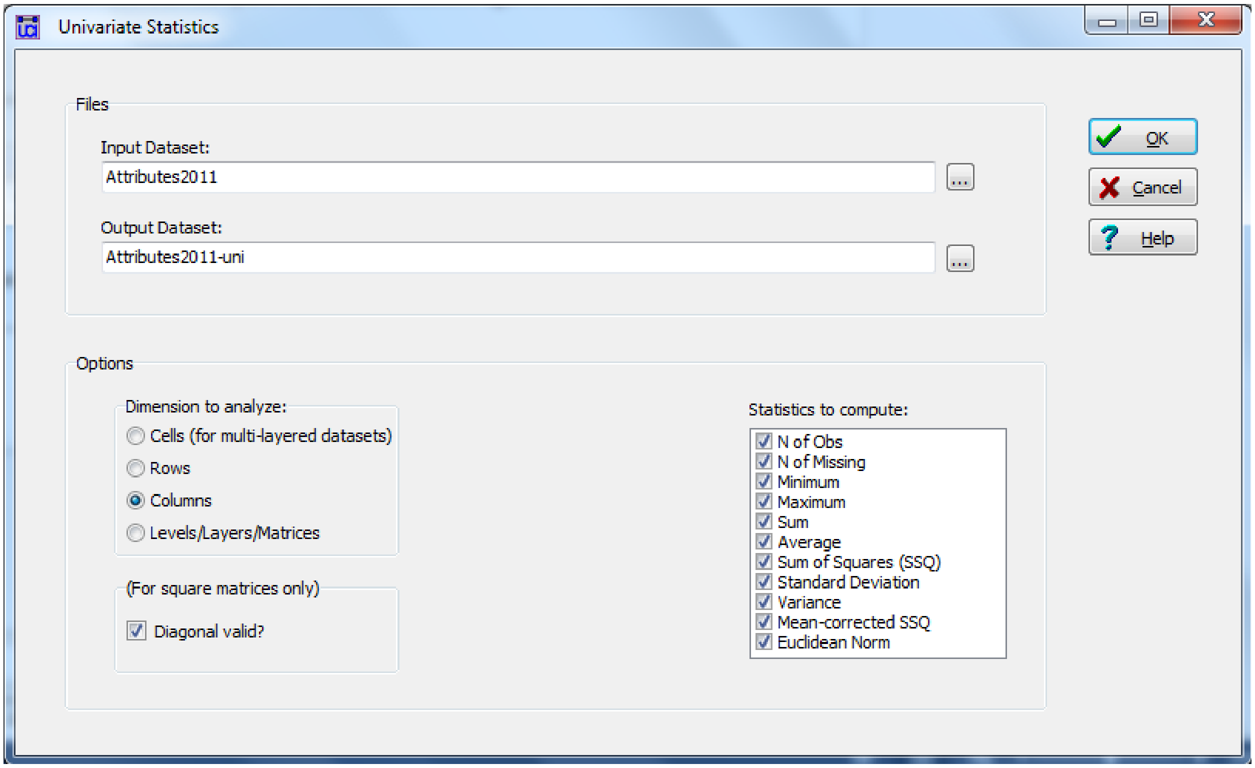
The basic idea of the permutation trials method is to take the existing data on the attributes of the nodes and randomly re-assign the scores on one attribute. The parameter we're interested in (for example, the Pearson correlation between student's attendance and test scores) is then measured in the permuted dataset, where the relationship between the two attributes is the result of a random trial. This procedure is repeated a number of times (say 1,000 or 10,000), and the distribution of the statistic across the random trials is calculated. We then compare the observed statistic against this random distribution to find out how frequently the statistic would be observed in random trials. It is a simple (if computationally a bit demanding) approach that preserves the observed distributions of both attributes – whatever they might happen to be – and requires no assumptions about sampling.

3.2 Univariate Statistics for Attributes

Usually the most interesting research questions call for examining association, partial association, and prediction. But, it is always a good idea to first examine each of the variables, one at a time. Since attribute data are stored as conventional rectangular data (cases by variables), any statistical package can be used to examine the distributions of the variables.

For quantitative attributes, UCINET's *Tools>Univariate Statistics* will do the trick. Let's look at our student data. The dialog is shown as figure 3.1.

Figure 3.1. UCINET Dialog for Attribute Univariate Statistics



Univariate Statistics

Files

Input Dataset: Attributes2011

Output Dataset: Attributes2011-uni

Options

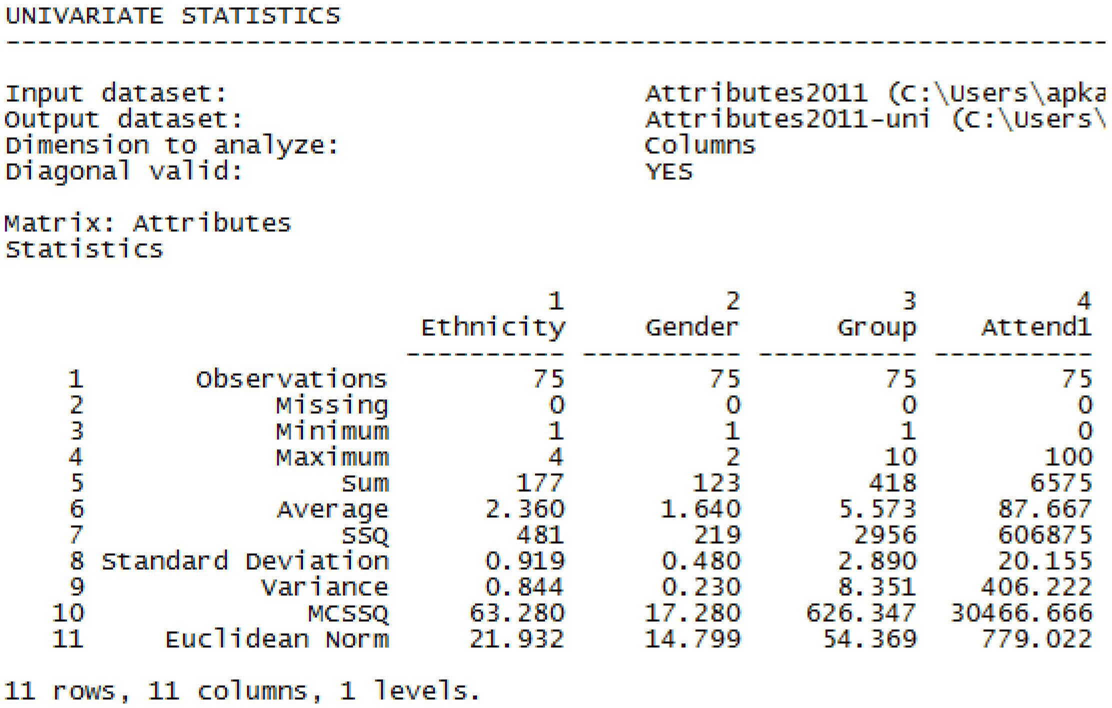
Dimension to analyze:
- ○ Cells (for multi-layered datasets)
- ○ Rows
- ● Columns
- ○ Levels/Layers/Matrices

(For square matrices only)
- ☑ Diagonal valid?

Statistics to compute:
- ☑ N of Obs
- ☑ N of Missing
- ☑ Minimum
- ☑ Maximum
- ☑ Sum
- ☑ Average
- ☑ Sum of Squares (SSQ)
- ☑ Standard Deviation
- ☑ Variance
- ☑ Mean-corrected SSQ
- ☑ Euclidean Norm

The dialog is very simple. We select the name of our attribute file, and specify "columns" to calculate the statistics on the distribution of the attributes across cases. The matrix is not node-by-node, so the question about the diagonal doesn't apply. The output can be saved as a UCINET data file (also rectangular, variables by statistics), and/or cut-and-pasted from the output log. A portion of the output is shown in figure 3.2.

Figure 3.2. Portion of UCINET Output for Attribute Univariate Statistics



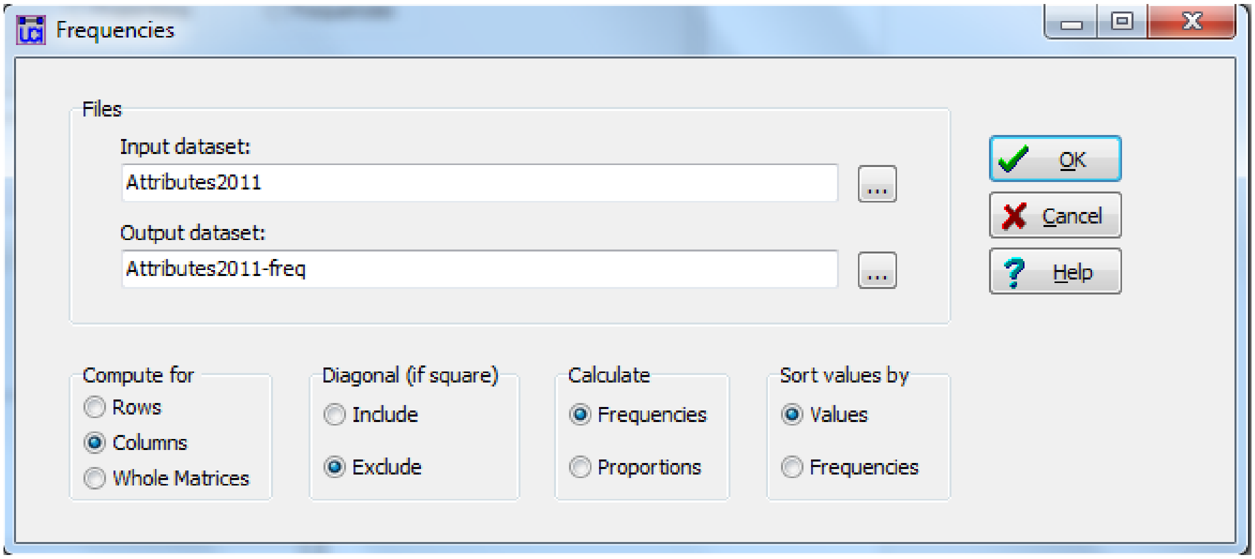
```
UNIVARIATE STATISTICS
--------------------------------------------------------------------

Input dataset:                          Attributes2011 (C:\Users\apka
Output dataset:                         Attributes2011-uni (C:\Users\
Dimension to analyze:                   Columns
Diagonal valid:                         YES

Matrix: Attributes
Statistics

                                 1          2          3          4
                         Ethnicity     Gender      Group    Attend1
                        ---------- ---------- ---------- ----------
  1        Observations         75         75         75         75
  2             Missing          0          0          0          0
  3             Minimum          1          1          1          0
  4             Maximum          4          2         10        100
  5                 Sum        177        123        418       6575
  6             Average      2.360      1.640      5.573     87.667
  7                 SSQ        481        219       2956     606875
  8  Standard Deviation      0.919      0.480      2.890     20.155
  9            Variance      0.844      0.230      8.351    406.222
 10               MCSSQ     63.280     17.280    626.347  30466.666
 11      Euclidean Norm     21.932     14.799     54.369    779.022

11 rows, 11 columns, 1 levels.
```

The report gives all the conventional basic descriptive statistics, N, minimum and maximum, as well as some additional measures of variation. UCINET doesn't compute skewness and kurtosis.

For our categorical variables of gender, ethnicity, and work group, moments are not particularly helpful. However, for continuous variables like Attend1 (percentage of in-class quizzes completed in weeks 1-3; up to the first exam), these statistics are useful.

A frequencies table would be nice, and UCINET does provide one – though it is not ideal. Figure 3.3 shows the dialog for *Tools>Frequencies*.

Figure 3.3. UCINET Dialog for Attribute Frequencies



Input dataset: Attributes2011

Output dataset: Attributes2011-freq

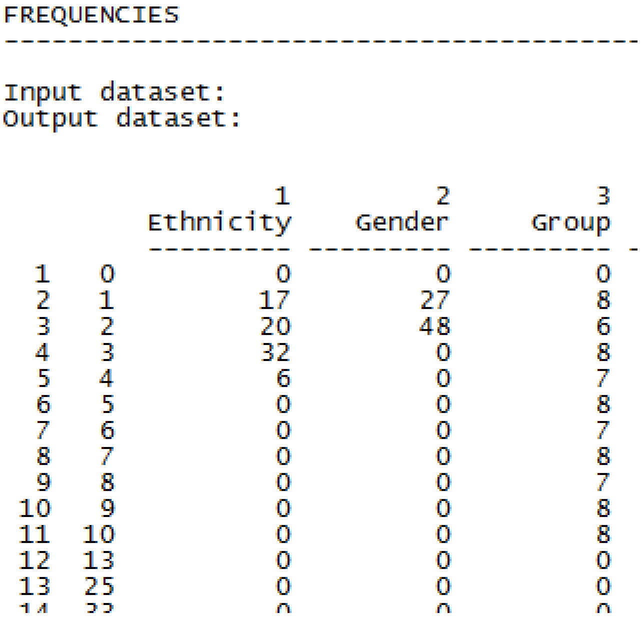
Compute for: ○ Rows ● Columns ○ Whole Matrices

Diagonal (if square): ○ Include ● Exclude

Calculate: ● Frequencies ○ Proportions

Sort values by: ● Values ○ Frequencies

Figure 3.4. Partial UCINET Output for Attribute Frequencies



```
FREQUENCIES
---------------------------------------

Input dataset:
Output dataset:


                  1         2         3
          Ethnicity    Gender     Group
          --------- --------- --------- -
 1    0           0         0         0
 2    1          17        27         8
 3    2          20        48         6
 4    3          32         0         8
 5    4           6         0         7
 6    5           0         0         8
 7    6           0         0         7
 8    7           0         0         8
 9    8           0         0         7
10    9           0         0         8
11   10           0         0         8
12   13           0         0         0
13   25           0         0         0
```

Figure 3.4 shows a portion of the output. We can still see that women outnumber men about 2 to 1 (48 to 27), that 17 students identified as White, 20 as Hispanic, 32 as Asian, and 6 as African American in the class.

Let's now turn to bi-variate association of nodal attributes.

3.3 Association Between the Attributes of Embedded Nodes

There are many different approaches to examining association or covariation between two variables, depending on the levels of measurement of the variables and the purposes of the analysis. Some of the most common approaches are to build cross-tabulations (for two categorical attributes), compare two or multiple group means (for one categorical and one continuous attribute), or use correlation and regression to examine two continuous attributes. UCINET does not have a tool for cross-tabulations and permutation tests for categorical association. For problems of that type, statistical software that will run boot-strapping, jackknife, or permutation should be used. In Stata, for example, one can embed a regular call for a tabular analysis within permutation trials with syntax such as:

Permute Y "text of the cross-tabulation command, options" name-of-statistic desired , reps(n)

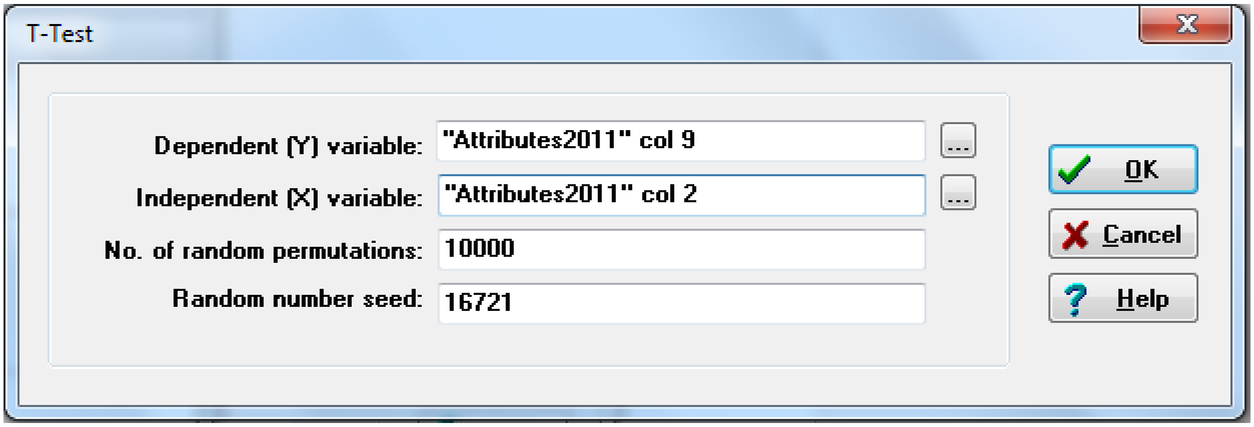
The permute command tells Stata to randomly permute the values of the variable Y (i.e. one of the variables in your cross-tab). The text of the tables command is then embedded in quotes (e.g. "tabulate Y X, chi2"). Following the crosstab command, a report on the values of one or more saved statistics is requested (e.g. perhaps chi-squared; saved in local memory as "r(chi2)" by the "tabulate" command), and the number of desired random replications is specified.

3.3.1 Comparing Two Groups

To test hypotheses about the association between a binary attribute (e.g. gender, in our data set) and a continuous one (e.g. exam scores), we can calculate a standard two-group t-test. In UCINET, the dialog can be located at: *Tools>Testing hypotheses>Node level>T-test.*

Let's look at two examples. We first look at whether there are gender differences (a fixed binary attribute) in mean final exam scores (a continuous attribute). Following conventional wisdom, our research hypothesis is one-tailed: we expect that the women's mean will be higher than the men's. Figure 3.5 shows the dialog.

Figure 3.5. UCINET Dialog for a Two-group T-Test



T-Test

Dependent (Y) variable: "Attributes2011" col 9

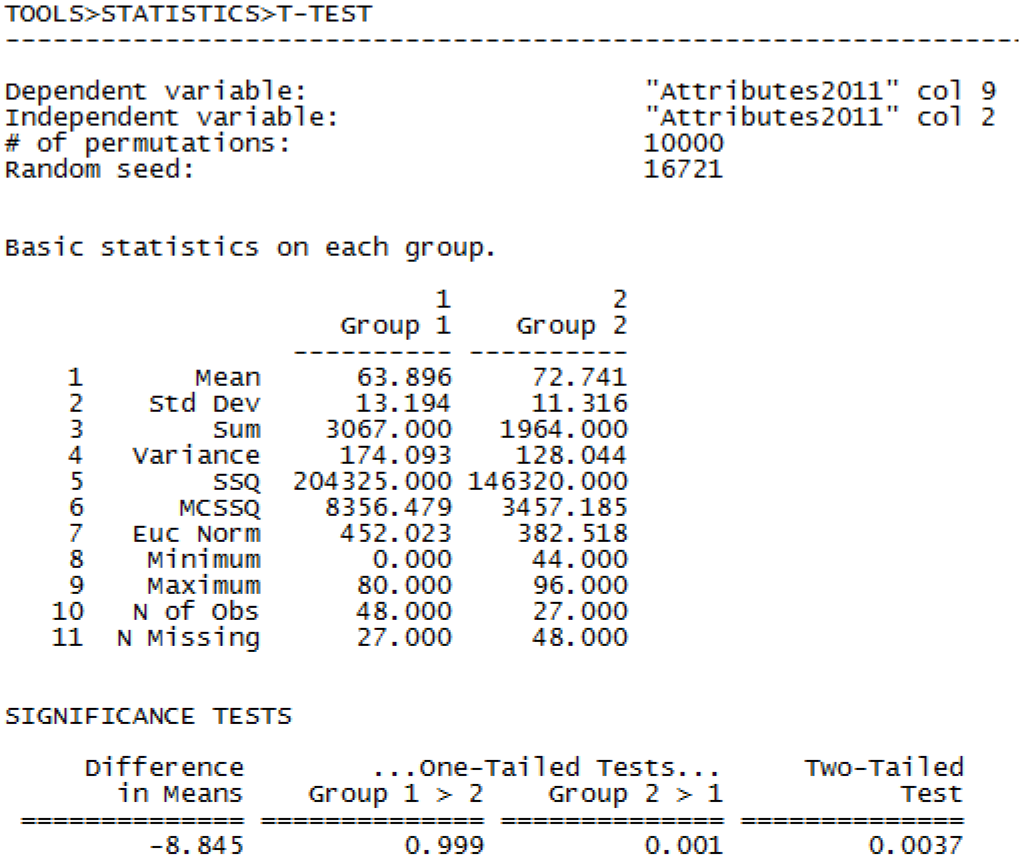
Independent (X) variable: "Attributes2011" col 2

No. of random permutations: 10000

Random number seed: 16721

Only two things to note here: First, you must know which column your variable is located in (final exam score happens to be column 9 in the attribute dataset, and gender happens to be column 2). UCINET will suggest column 1 by default, and you can simply edit it. Second, you can select the number of random permutations for significance tests. The default is 10,000, which is more than adequate for most purposes. Note the "random number seed." You may wish to generate *the same* permuted distribution for multiple tests. This can be done by noting, and then specifying that the same seed be used to start the pseudo-random number generator. Figure 3.6 shows the output (try replicating Figure 3.6 using the seed 16721).

Figure 3.6. UCINET Output for Testing Gender Differences in Final Exam Scores

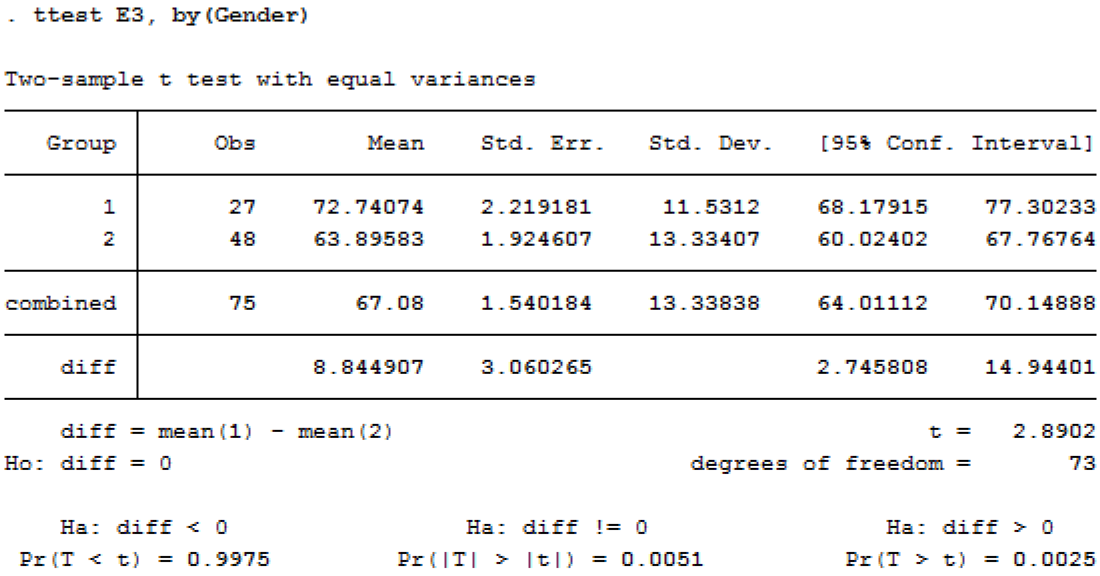


```
TOOLS>STATISTICS>T-TEST
----------------------------------------------------------------

Dependent variable:                     "Attributes2011" col 9
Independent variable:                   "Attributes2011" col 2
# of permutations:                      10000
Random seed:                            16721


Basic statistics on each group.

                                   1           2
                             Group 1     Group 2
                          ---------- ----------
   1        Mean          63.896      72.741
   2     Std Dev          13.194      11.316
   3         Sum        3067.000    1964.000
   4    Variance         174.093     128.044
   5         SSQ      204325.000  146320.000
   6       MCSSQ        8356.479    3457.185
   7    Euc Norm         452.023     382.518
   8     Minimum           0.000      44.000
   9     Maximum          80.000      96.000
  10    N of Obs          48.000      27.000
  11   N Missing          27.000      48.000


SIGNIFICANCE TESTS

       Difference       ...One-Tailed Tests...      Two-Tailed
         in Means      Group 1 > 2    Group 2 > 1         Test
  ============== ============== ============== ==============
           -8.845          0.999          0.001         0.0037
```

Here, we need to note that UCINET takes the value of the categorical variable (gender) from the first case and calls that "Group 1". So in this output, Group 1 happens to be women (coded 2 on the variable) due to the arbitrary fact that the first case in the dataset is a woman. This can be confusing, so the user should always check the attributes of the first node (or the "N" value if they differ; for example, we know the N for women is 48 from figure 3.4, therefore Group 1 in Figure 3.6 is women) to make sure they know which group is which. The output shows the descriptive statistics for the two groups. Women have a mean test score of about 64, while the men's average is about 73, in contradiction to our research hypothesis. The variation within each group, however, is substantial, and there are considerably more women than men. In the significance tests section, we see the difference in means and three probability levels based on the standard error of the difference between means generated from the permutation trials. The interpretation is as follows: in over 99.9%

of random networks with the same numbers of men and women and the same univariate distribution of test scores, the mean of group 1 (women) is NOT higher than the mean of group 2 (men). In more than 99.9% of the trials, the mean of group 2 (men) is higher than group 1 (women). In about one third of 1% of the permutation trials, the difference in means observed between groups is less than the average difference observed in random trials.

Figure 3.7. Stata Test of Gender Difference in Final Exam Scores with Conventional Standard Errors

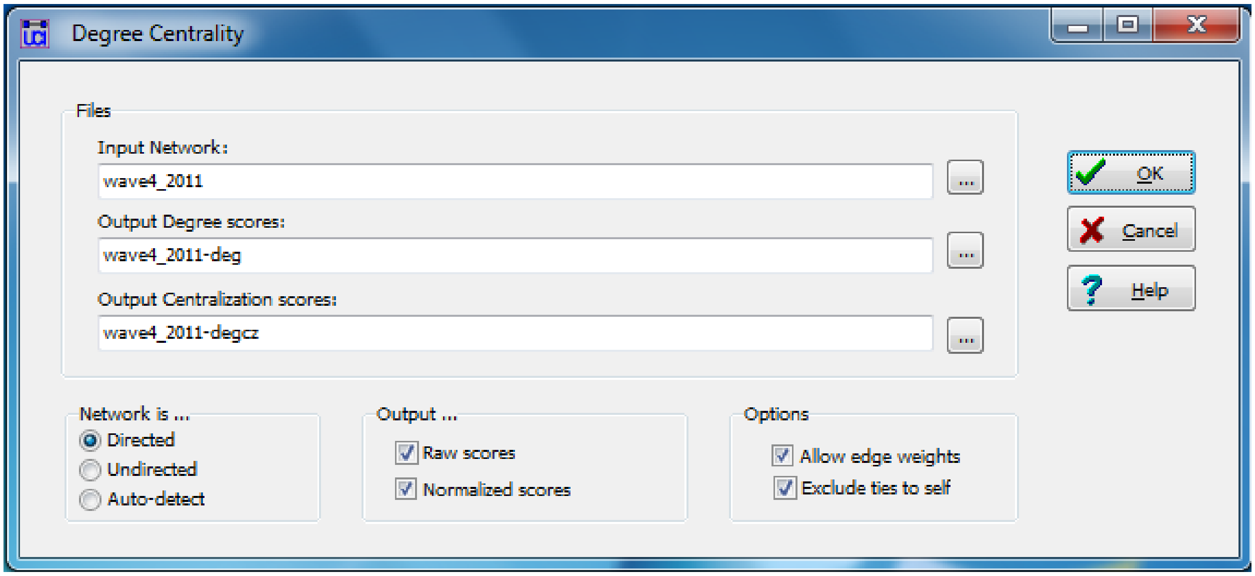


```
. ttest E3, by(Gender)

Two-sample t test with equal variances
------------------------------------------------------------------------------
   Group |     Obs        Mean    Std. Err.   Std. Dev.   [95% Conf. Interval]
---------+--------------------------------------------------------------------
       1 |      27    72.74074    2.219181     11.5312    68.17915    77.30233
       2 |      48    63.89583    1.924607    13.33407    60.02402    67.76764
---------+--------------------------------------------------------------------
combined |      75       67.08    1.540184    13.33838    64.01112    70.14888
---------+--------------------------------------------------------------------
    diff |            8.844907    3.060265                2.745808    14.94401
------------------------------------------------------------------------------
    diff = mean(1) - mean(2)                                      t =   2.8902
Ho: diff = 0                                     degrees of freedom =       73

    Ha: diff < 0                 Ha: diff != 0                 Ha: diff > 0
 Pr(T < t) = 0.9975         Pr(|T| > |t|) = 0.0051          Pr(T > t) = 0.0025
```

Figure 3.7 shows the same problem, but tested in Stata using conventional standard errors. In this case, we see that the statistical significance (p level) of the test of the difference between the two group's means is stronger using the standard error estimated by permutation trials than using the classical formula (the two-tail p-level for the permutation test is 0.0037, the two-tail p-level using the classical formula is 0.0051). Standard errors and significance results may be either stronger or weaker using the permutation method than the classical formulas.

In the next example, we ask whether the average in-degree of women (that is, the number of others who say they are acquainted with them) is different from the average in-degree of men at the time of the final exam. We can calculate the in-degree of each node using *Network>Centrality and Power> Degree* in UCINET.

Figure 3.8. UCINET Dialog for Generating Degree Data



Files

Input Network: wave4_2011

Output Degree scores: wave4_2011-deg

Output Centralization scores: wave4_2011-degcz

Network is ...

● Directed
○ Undirected
○ Auto-detect

Output ...

☑ Raw scores
☑ Normalized scores

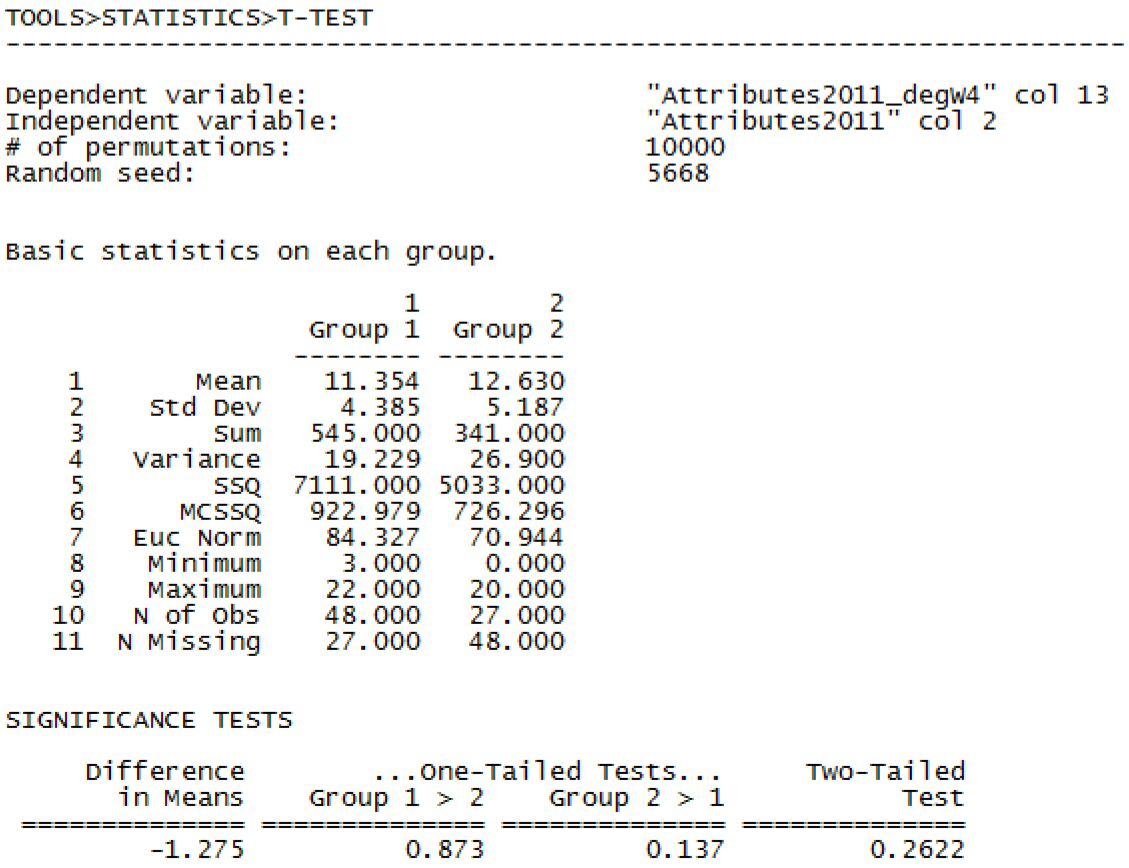
Options

☑ Allow edge weights
☑ Exclude ties to self

Here, we've chosen the Wave 4 relational data, specified that it's directed, and UCINET saves measures of degree and centrality in separate files. We can use the *Data>Join>Join Columns* procedure to append the degree attributes generated in the file "wave4_2011-deg" to our "Attributes2011" file and create a new dataset called "Attributes2011_degW4" (the *Join* procedure was discussed in greater detail in Chapter 2). When running a t-test for differences in in-degree between men and women, the dialog is the same as in Figure 3.5, except that we have selected different variables (in-degree will be column 13 in the new joined dataset).

The reason for a second example of a simple two-group t-test is to point out that attributes can be measures of how individuals are embedded in social networks. The in-degree of a node describes dyadic relations (someone else said they were acquainted with ego). But, the extent to which a node has dyadic ties is an attribute of the node itself. The output from the t-test is shown in Figure 3.9.

Figure 3.9. UCINET Difference in In-degree Between Men and Women Students (Wave 4)



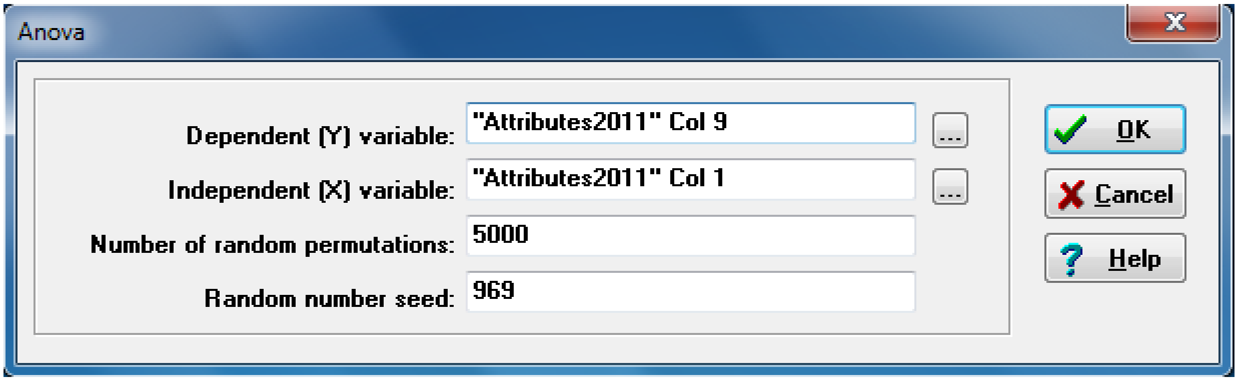
```
TOOLS>STATISTICS>T-TEST
------------------------------------------------------------------------

Dependent variable:                     "Attributes2011_degW4" col 13
Independent variable:                   "Attributes2011" col 2
# of permutations:                      10000
Random seed:                            5668


Basic statistics on each group.

                           1        2
                     Group 1  Group 2
                    -------- --------
   1        Mean      11.354   12.630
   2     Std Dev       4.385    5.187
   3         Sum     545.000  341.000
   4    Variance      19.229   26.900
   5         SSQ    7111.000 5033.000
   6       MCSSQ     922.979  726.296
   7    Euc Norm      84.327   70.944
   8     Minimum       3.000    0.000
   9     Maximum      22.000   20.000
  10    N of Obs      48.000   27.000
  11   N Missing      27.000   48.000


SIGNIFICANCE TESTS

     Difference        ...One-Tailed Tests...     Two-Tailed
       in Means     Group 1 > 2    Group 2 > 1          Test
 ============== ============== ============== ==============
         -1.275          0.873          0.137         0.2622
```

We see that the mean in-degree of men students (group 2) at the end of the course was a bit higher than that of women. Using the standard errors of the difference in means generated by permutation tests, we see that a difference this large occurs relatively frequently in randomly permuted networks (26% of the random networks). Therefore, we do not find support for the notion that the difference in the average in-degree of men and women students is not due to random processes.

3.3.2 Comparing Multiple Groups

If one of the attributes is categorical with more than two categories, and the other attribute is continuous, a common approach to testing hypotheses is one-way ANOVA (in UCINET: *Tools>Testing Hypotheses>Node Level>ANOVA*). Let's see if there are any significant differences between the mean final exam scores of students classified by ethnicity. The dialog is in figure 3.10, and the output is in figure 3.11.

Figure 3.10. UCINET Final Exam Score Means by Ethnicity Dialog



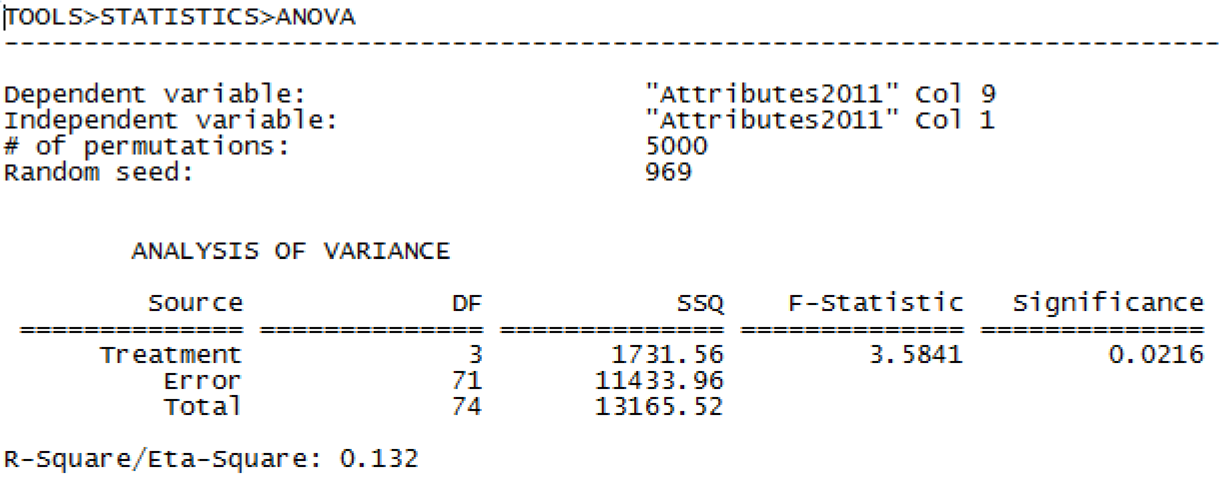
Dependent (Y) variable: "Attributes2011" Col 9

Independent (X) variable: "Attributes2011" Col 1

Number of random permutations: 5000

Random number seed: 969

Figure 3.11. UCINET Final Exam Score Means by Ethnicity Output

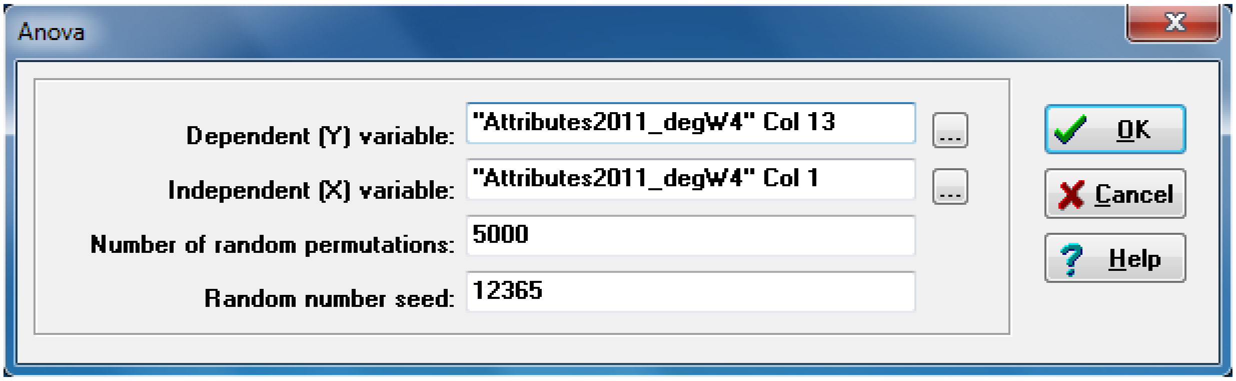


```
TOOLS>STATISTICS>ANOVA
------------------------------------------------------------------------------

Dependent variable:                     "Attributes2011" Col 9
Independent variable:                   "Attributes2011" Col 1
# of permutations:                      5000
Random seed:                            969


        ANALYSIS OF VARIANCE

         Source              DF            SSQ   F-Statistic  Significance
 ============== ============== ============== ============== ==============
      Treatment               3        1731.56        3.5841         0.0216
          Error              71       11433.96
          Total              74       13165.52

R-Square/Eta-Square: 0.132
```

UCINET's output is minimal, providing the standard ANOVA table, F-test statistic, and significance. Eta-square is also shown. We conclude that to 95% confidence, there is at least one difference between group means that is not expected to occur very frequently in random permutations of the data. And, group mean differences account for 13.2% of the observed variation in individual's final exam scores. That is, the differences among mean final exam scores by ethnic identity probably are not due to random variation (though there may very well be spurious factors at work here).

Let's see whether there are differences by ethnic identity in the extent to which individuals are named by others as acquaintances.

Figure 3.12. UCINET Node In-degree by Ethnicity Dialog



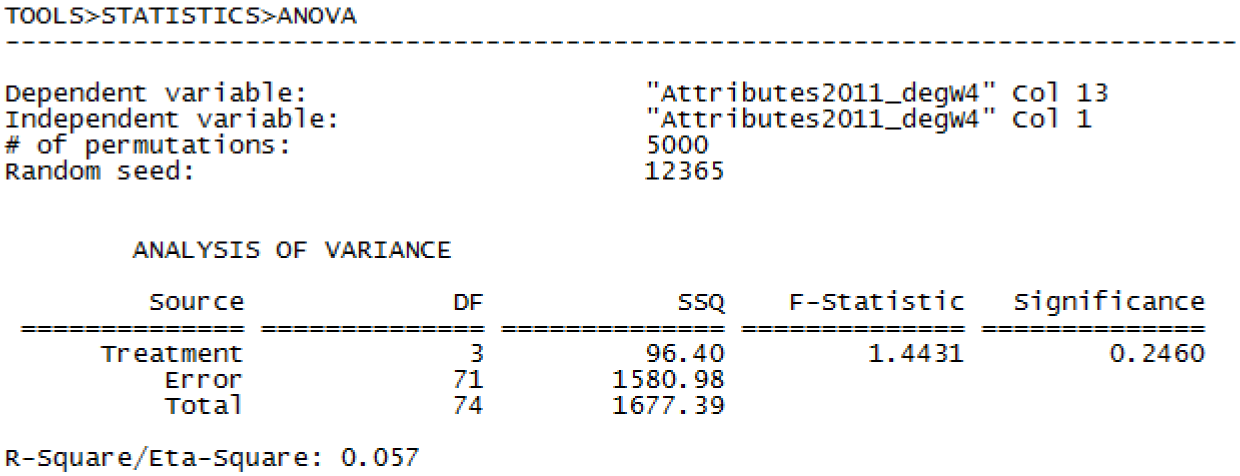
Dependent (Y) variable: "Attributes2011_degW4" Col 13

Independent (X) variable: "Attributes2011_degW4" Col 1

Number of random permutations: 5000

Random number seed: 12365

Figure 3.13. UCINET Node In-degree by Ethnicity Output



```
TOOLS>STATISTICS>ANOVA
------------------------------------------------------------------------------

Dependent variable:                      "Attributes2011_degW4" Col 13
Independent variable:                    "Attributes2011_degW4" Col 1
# of permutations:                       5000
Random seed:                             12365


        ANALYSIS OF VARIANCE

         Source             DF            SSQ    F-Statistic   Significance
 ============== ============== ============== ============== ==============
      Treatment              3          96.40         1.4431         0.2460
          Error             71        1580.98
          Total             74        1677.39

R-Square/Eta-Square: 0.057
```

In the classroom data, it looks like there are no reliable differences due to ethnicity in the extent to which students are likely to be known by others in the class.

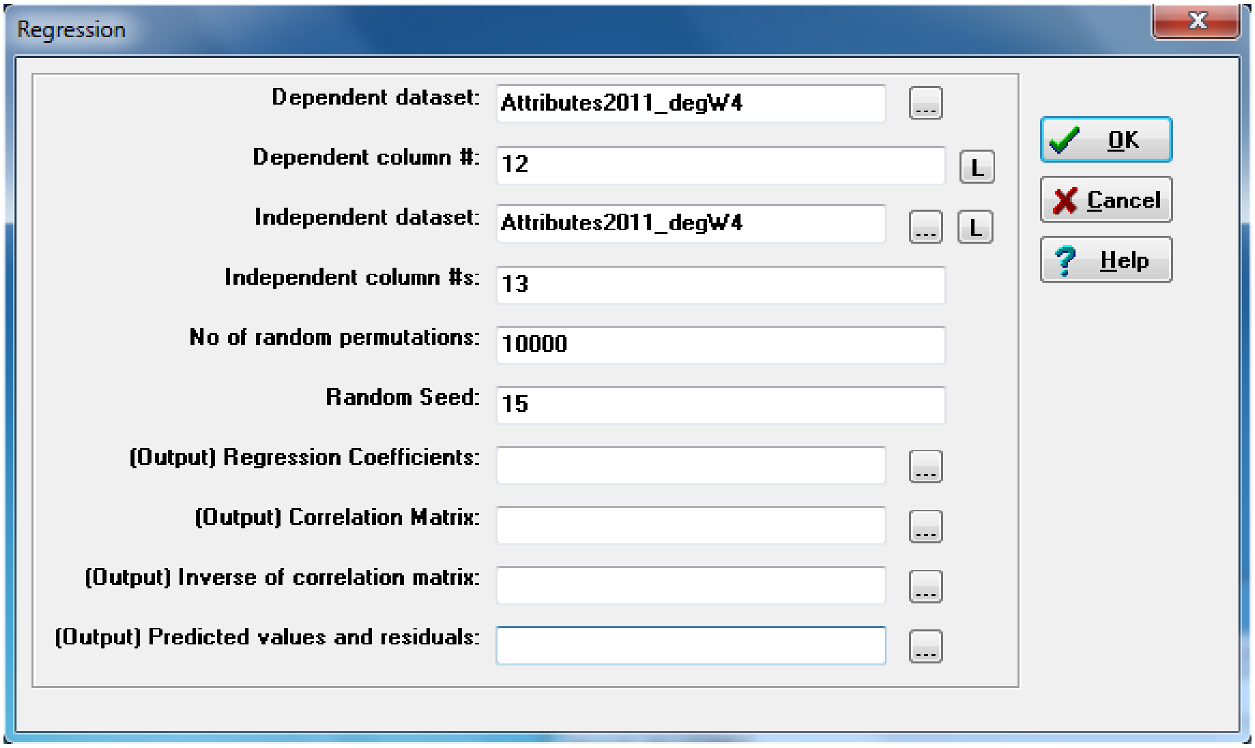
3.3.3 Continuous Association

Where one or both attributes are measured at the ordinal level, the standard approach is to calculate measures of association (e.g. gamma). UCINET doesn't have built in tools for calculating hypothesis tests for grouped-ordinal variables. Statistical software packages that allow calculation of estimated standard errors with re-sampling or permutation should be used.

For problems where both attributes are interval-ratio, or can reasonably be treated as such, UCINET has built-in tools for using linear regression (*Tools>Testing hypotheses>Regression*).

Let's examine whether students who name more others as acquaintances than average (out-degree) are, themselves, more frequently cited by others (in-degree). Again, variables that describe how individuals are embedded in the network (in-degree, out-degree) are being treated as attributes of the individuals. Obviously, standard significance tests don't apply, as the nodes generating the degree counts are the same individuals.

Figure 3.14. UCINET In-degree and Out-degree Regression Dialog



Regression

Dependent dataset: Attributes2011_degW4

Dependent column #: 12

Independent dataset: Attributes2011_degW4

Independent column #s: 13

No of random permutations: 10000

Random Seed: 15

(Output) Regression Coefficients:

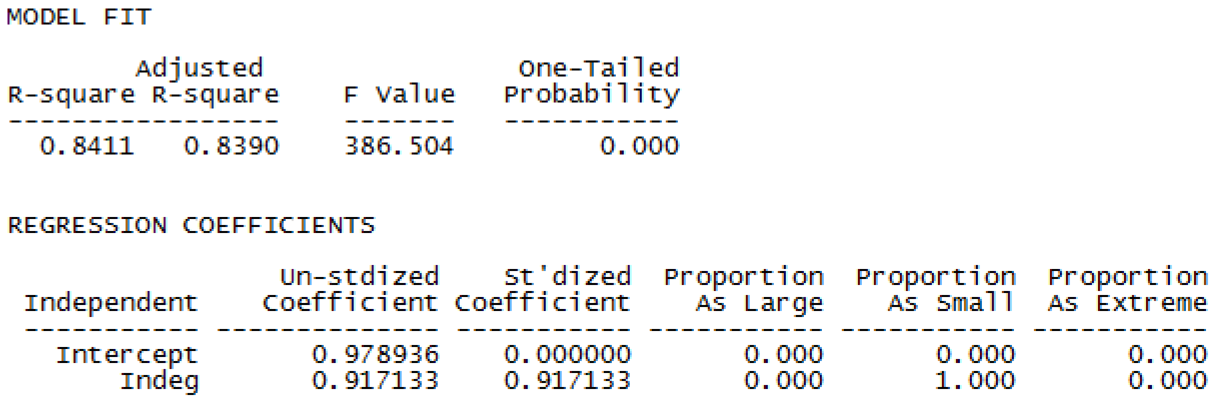
(Output) Correlation Matrix:

(Output) Inverse of correlation matrix:

(Output) Predicted values and residuals:

In the dialog, note that the dependent and independent variable can be taken from different datasets (one might be a file of attributes, the other a file of results of calculating network statistics like degree or centrality). All of the important regression output can be saved in output data files for further processing. We've left the boxes blank here so as not to save the output as new datasets.

Figure 3.15. Portion of UCINET In-degree and Out-degree Linear Regression Output

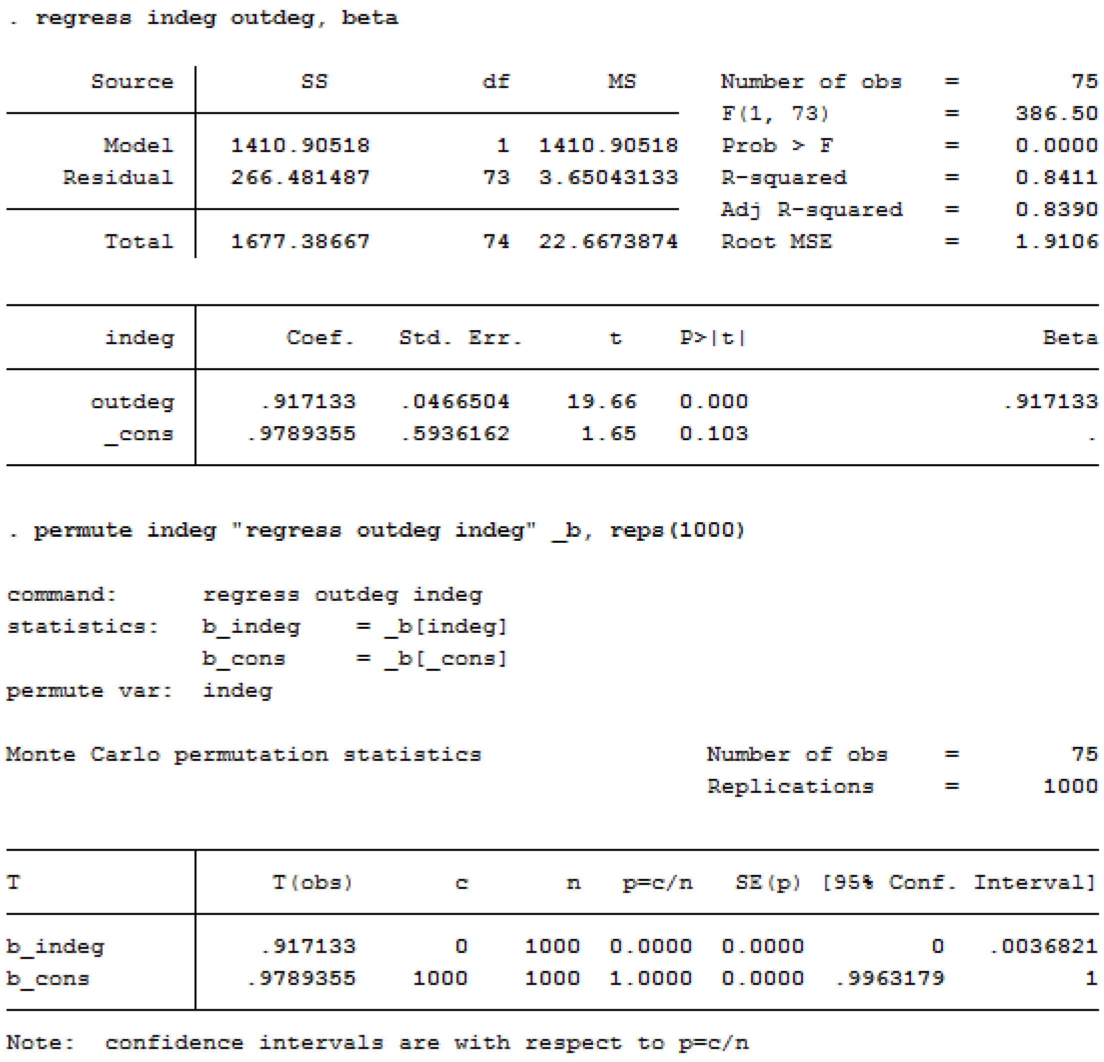


```
MODEL FIT

        Adjusted                  One-Tailed
R-square R-square    F Value     Probability
-----------------    -------     -----------
  0.8411   0.8390    386.504           0.000

REGRESSION COEFFICIENTS

                      Un-stdized     St'dized  Proportion  Proportion  Proportion
 Independent     Coefficient Coefficient    As Large    As Small  As Extreme
 -----------  --------------- -----------  -----------  -----------  -----------
   Intercept         0.978936    0.000000       0.000       0.000       0.000
       Indeg         0.917133    0.917133       0.000       1.000       0.000
```

From the output, we see that there is a strong positive association on out-degree with in-degree (+0.917). The variance explained in in-degree by out-degree is considerable (0.841). Each additional person named by ego as an acquaintance is associated with an increase of 0.917 others naming ego as an acquaintance. A two-tailed p-level for the coefficient is reported as < 0.001 (under "Proportion As Extreme").

Let's run the same problem in Stata, using both the conventional standard errors and permutation trials to test the slope coefficient.

Figure 3.16. Stata Regression Output with and without Permutation Trials



```
. regress indeg outdeg, beta

      Source |       SS       df       MS              Number of obs =      75
-------------+------------------------------           F(1, 73)      =  386.50
       Model |  1410.90518     1  1410.90518           Prob > F      =  0.0000
    Residual |  266.481487    73  3.65043133           R-squared     =  0.8411
-------------+------------------------------           Adj R-squared =  0.8390
       Total |  1677.38667    74  22.6673874           Root MSE      =  1.9106

------------------------------------------------------------------------------
       indeg |      Coef.   Std. Err.      t    P>|t|                     Beta
-------------+----------------------------------------------------------------
      outdeg |    .917133   .0466504    19.66   0.000                  .917133
       _cons |   .9789355   .5936162     1.65   0.103                        .
------------------------------------------------------------------------------

. permute indeg "regress outdeg indeg" _b, reps(1000)

command:      regress outdeg indeg
statistics:   b_indeg    = _b[indeg]
              b_cons     = _b[_cons]
permute var:  indeg

Monte Carlo permutation statistics                Number of obs    =      75
                                                  Replications     =    1000

------------------------------------------------------------------------------
T            |     T(obs)       c       n   p=c/n   SE(p) [95% Conf. Interval]
-------------+----------------------------------------------------------------
b_indeg      |    .917133       0    1000  0.0000  0.0000         0   .0036821
b_cons       |   .9789355    1000    1000  1.0000  0.0000  .9963179          1
------------------------------------------------------------------------------
Note:  confidence intervals are with respect to p=c/n
```

Stata reproduces the regression coefficients and R^2 statistics. Notice that the conventional standard error approach and permutation trial approach produce identical results for this example.

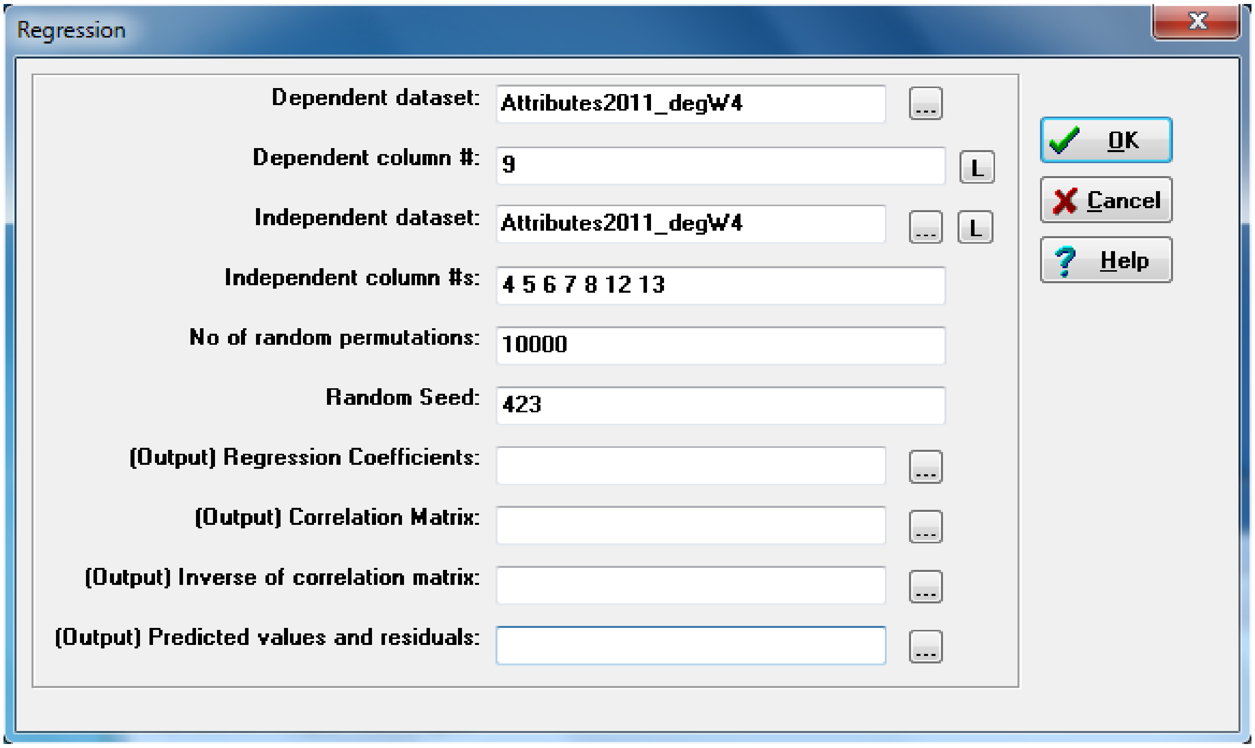
3.4 Partial Association and Prediction of Attributes by Attributes

While many useful hypotheses can be addressed with simple bi-variate association, most of our work uses variations of the generalized linear model to implement statistical control via partialling. UCINET has basic tools for multiple linear regression with permutation tests, and this is a nice tool when the attribute we are interested in predicting is measured at the interval-ratio level. Because the significance tests are based on permutations, we do not need to assume the normality of the distribution of residuals.

3.4.1 Multiple Regression

As an example, let's extend our efforts to predict which students were more likely to perform well on the final exam (E3). Here's an example of the UCINET dialog for regression that specifies multiple independent variables.

3.17. UCINET Multiple Regression Dialog



Regression

Dependent dataset: Attributes2011_degW4

Dependent column #: 9

Independent dataset: Attributes2011_degW4

Independent column #s: 4 5 6 7 8 12 13

No of random permutations: 10000

Random Seed: 423

(Output) Regression Coefficients:

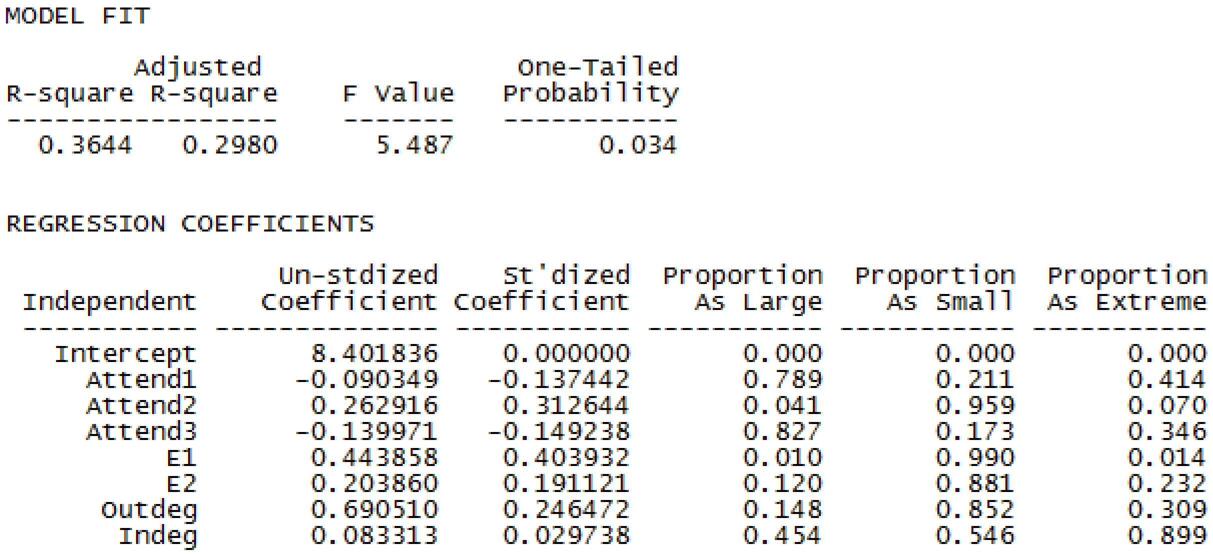
(Output) Correlation Matrix:

(Output) Inverse of correlation matrix:

(Output) Predicted values and residuals:

The dependent variable is selected as the desired column in one attribute data set. The independent variables are selected by column numbers from the same data set, or a single different data set. The number of permutations and seed can be selected, and all of the basic regression output components can be saved for further processing or reporting. The output is presented in Figure 3.18.

3.18. Portion of UCINET Multiple Regression Output

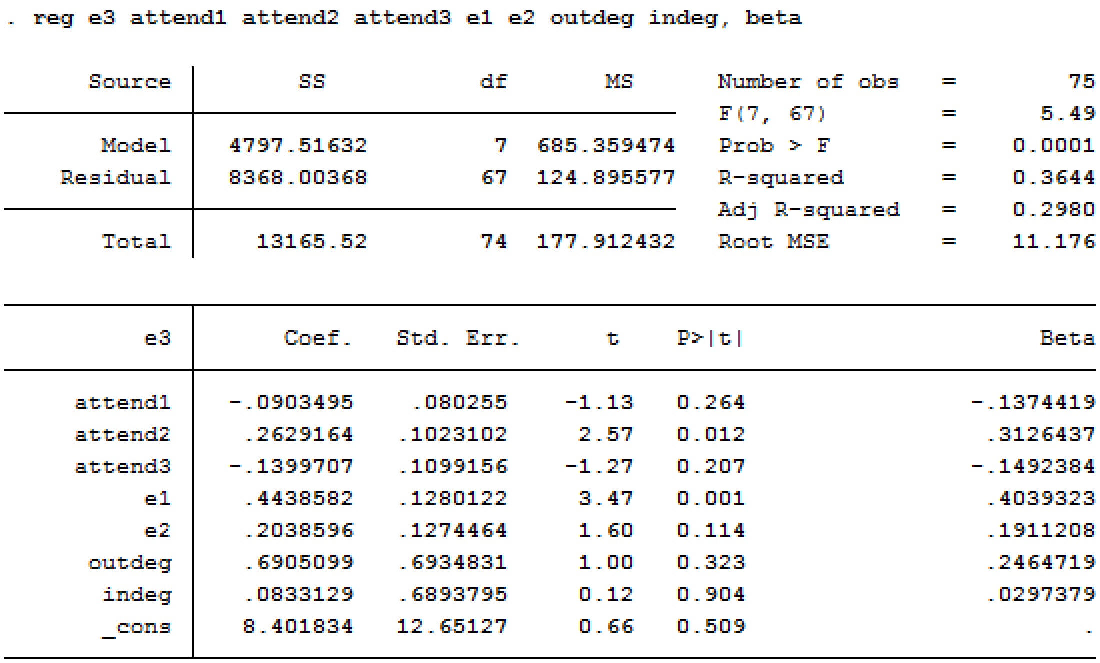


```
MODEL FIT

         Adjusted                  One-Tailed
R-square R-square    F Value   Probability
-----------------    -------   -----------
  0.3644   0.2980      5.487         0.034

REGRESSION COEFFICIENTS

                  Un-stdized    St'dized  Proportion  Proportion  Proportion
 Independent     Coefficient Coefficient    As Large    As Small  As Extreme
 ----------- --------------- ----------- ----------- ----------- -----------
   Intercept        8.401836    0.000000       0.000       0.000       0.000
     Attend1       -0.090349   -0.137442       0.789       0.211       0.414
     Attend2        0.262916    0.312644       0.041       0.959       0.070
     Attend3       -0.139971   -0.149238       0.827       0.173       0.346
          E1        0.443858    0.403932       0.010       0.990       0.014
          E2        0.203860    0.191121       0.120       0.881       0.232
      Outdeg        0.690510    0.246472       0.148       0.852       0.309
       Indeg        0.083313    0.029738       0.454       0.546       0.899
```

Scores on the final exam (E3) turn out to be moderately predictable, with a significant F-value and adjusted R-square of about 0.3. Score on the first midterm (E1) has the largest significant effect on the final exam score, but attendance during the middle of the term is also a significant predictor (p < 0.05, one-tail). Score on the second midterm (E2) also has a positive effect on E3, but only in 88% of random trials which cannot be considered significant based on typical social science statistical standards. Neither measure of embeddedness is strongly related to exam performance, with nodes that are more popular (in-degree) than we would expect for their sociability (out-degree) showing a tendency to perform better, net of other factors. But again, this effect is not significant to 95%.

Finally, let's compare the results of the permutation tests to the regular parametric tests, which are shown in Figure 3.19, where they were calculated with Stata.

3.19. Stata Regression Output



```
. reg e3 attend1 attend2 attend3 e1 e2 outdeg indeg, beta

      Source |       SS       df       MS              Number of obs =      75
-------------+------------------------------           F(7,  67)     =    5.49
       Model |  4797.51632     7  685.359474           Prob > F      =  0.0001
    Residual |  8368.00368    67  124.895577           R-squared     =  0.3644
-------------+------------------------------           Adj R-squared =  0.2980
       Total |    13165.52    74  177.912432           Root MSE      =  11.176

------------------------------------------------------------------------------
          e3 |      Coef.   Std. Err.      t    P>|t|                     Beta
-------------+----------------------------------------------------------------
     attend1 |  -.0903495    .080255    -1.13   0.264                -.1374419
     attend2 |   .2629164   .1023102     2.57   0.012                 .3126437
     attend3 |  -.1399707   .1099156    -1.27   0.207                -.1492384
          e1 |   .4438582   .1280122     3.47   0.001                 .4039323
          e2 |   .2038596   .1274464     1.60   0.114                 .1911208
      outdeg |   .6905099   .6934831     1.00   0.323                 .2464719
       indeg |   .0833129   .6893795     0.12   0.904                 .0297379
       _cons |   8.401834   12.65127     0.66   0.509                        .
------------------------------------------------------------------------------
```

More accurate estimates using permutation trials guard us against committing errors of assuming that effects are systematic, when they might well really be unreliable.

3.4.2 Generalized Linear Models

Where attributes of interest are not Gaussian, one should use a statistical package with GLM and appropriate dependent distribution and link functions. Most scientific statistical software suites include a wide variety of modeling approaches for attributes with varying properties. Generally, Monte Carlo or permutation trials methods are available to get corrected standard errors and significance tests – though sometimes it requires a bit of work.

Attributes of nodes might sometimes be a count of something. For example, we might count the number of children that an ego has had and explore whether this is associated with the attributes of ego's friends. A Poisson or Negative Binomial model might be used.

Attributes of nodes might be binary (for example, ego is a drug user, or not). Logistic, probit, or complementary log-log models might be used.

Attributes of nodes might be “multiple choice” outcomes (e.g. ego is working full time, part time, unemployed, or not in the labor force). Multinomial logits or probits might apply, or ordinal cumulative logits and probits. See textbooks describing general linear modeling for details (e.g. Hoffman, 2004).

3.5 Summary

Virtually any hypothesis about the relationships between nodal attributes in SNA data can be studied using the modeling techniques that are used for data based on other observational schemes.

Social network data, though, are usually populations rather than samples. So, the questions of “generalization” from the sample to the population usually do not arise. Inferential statistics are necessary, though, as the results in one observation of a network may appear to be systematic, but are really the result of an unusual outcome of a random process. Permutation approaches (Monte Carlo simulation) provide a natural tool for testing the reliability of results with social network data.

Measures of network position can be used as attributes of individuals (for example, degree, centrality, closure of ego networks, homophily of ego net, proportion of others who have adopted or have some attribute, or the average score of those connected to ego). Conventional statistics then can be used to include some powerful social influence and embedding information in tests about ego’s attributes, in addition to ego’s fixed individual attributes. In a later chapter we will look at some other ways of studying how the attributes of those to whom a node is connected might impact the node’s attributes (network influence models are discussed in Chapter 6).

Conventional approaches to estimating standard errors and hypothesis testing are not appropriate when using social network data, because the “cases” (nodes) are not independent. But, this turns out to not be much of a problem. The logic of permutation trials to generate estimates of reliability of parameters fits naturally with SNA. UCINET adopts this approach for some common tests. Virtually any hypothesis though, can be tested by using permutation in combination with generalized linear models.

In this chapter, we've discussed how to study the relationship between two (or more) attributes of nodes (or, variables measured on cases). The unique contribution of SNA, however, lies in treating the relationship or tie between nodes or cases as the thing to be explained and understood. In the next chapter, we'll take a look at some approaches to studying the relationships between two or more attributes of relations, or dyads.

3.6 References

Efron, Bradley. 1981. "Nonparametric Estimates of Standard Error: The Jackknife, the Bootstrap and Other Methods," *Biometrika*, 68(3): 589-599.

Hoffman, John. 2004. *Generalized Linear Models: An Applied Approach*. Pearson.

Chapter 4. Association Between Networks

Relational data describe the ties between pairs of actors. Just as we might ask whether two attributes are associated (e.g. do men and women differ on test scores?), we might ask whether two relations are associated (e.g. are people who are connected by being in the same work group likely to be connected by a friendship tie?). This chapter introduces analyses that examine association between relational data.

4.1 Networks as Dyads

In the last chapter we looked at how the relationships among attributes of nodes in a network can be studied statistically. This is really the same as studying statistical association between variables, observed across cases. The only new issue was how to test hypotheses appropriately in the face of the non-independence of the cases.

In this chapter we take a look at studying association between networks. That is, are the relations of one type among the actors in a network associated with relations of another type? For example, are people who are friends outside of work more likely to go to one another for advice at work? In our example student data (introduced in Chapter 2), we might ask how similar the patterns of acquaintanceship are between the beginning and the end of the class, or whether people who were both in the same work group are more likely to be acquainted with each other.

Studying whether two or more networks are associated is actually quite straightforward. Rather than treating the "case" or node as a unit of observation (or a row in our dataset), we treat a "dyad" of nodes as the unit of observation. A first network is "deconstructed" into all possible pairs of nodes (dyads), and the dyadic relationship is measured for each pair (it may be binary, i.e. present/absent—or valued, i.e. tie strength). The other network of interest is deconstructed in the same way. Then, the association between the two can be calculated. The "sample size" then is always equal to the number of unique pairs of nodes. If the network is directed,

$$N = K(K-1) \qquad 4.1$$

where K is the number of nodes in the network. If the network is symmetric,

$$N = \frac{K(K-1)}{2} \qquad 4.2$$

The trick to studying associations among networks lies in seeing the relation between each pair of nodes as the object of interest (is a tie present?, how strong is it?). A network is seen as simply a collection of (all possible) dyads. The association between two networks is describing the extent to which the scores of one set of relations correspond to the scores of another set of relations among the same actors.

In studying the relationships among networks, the inferential statistical question is not one of generalizability – after all, we have (in theory) directly observed the entire population of actors, so there is no inference from sample to population. But, there is a question of how likely it is that the association we observe between two or more networks is the result of random, rather than systematic processes. So, again, the method of generating standard errors by permutation of the existing data is a useful tool.

Let's take a look at a few simple examples of the association and partial association among networks using tools from the UCINET toolkit.

4.2 Association Between Networks

Two networks are associated, co-vary, or are correlated to the extent that the patterns of dyadic relations in one correspond to the pattern of dyadic relations in the other. The relation between the actors in a dyad can be measured as either present/absent, or as a matter of degree. Let's start with categorical relations, which are studied by way of cross-tabulation, then turn to continuous relations, which are studied by way of correlation.

4.2.1 Categorical Relations

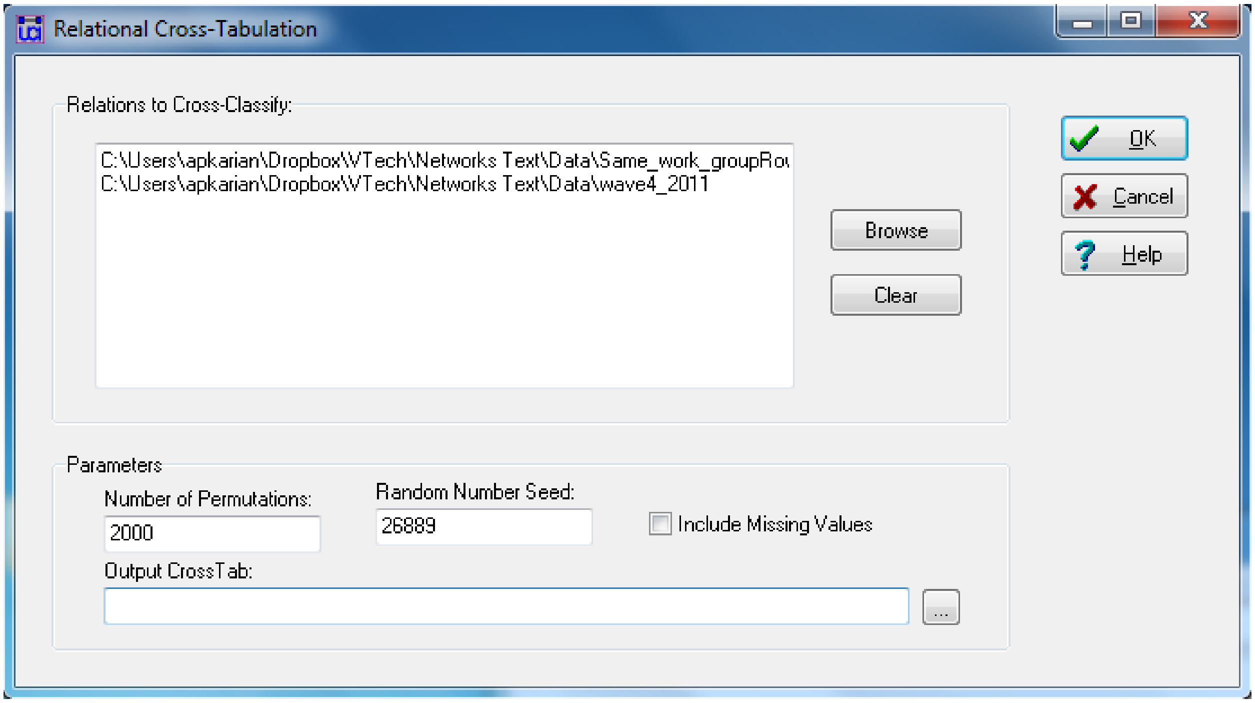
Suppose that we want to know whether there is an association between two categorical relations. This somewhat new idea is far easier to grasp with examples. Here's an obvious hypothesis about the association between two categorical relations: students who were assigned to the same groups to work on their term papers are more likely to report that they are acquainted by the end of the course than students who were not in the same workgroup.

One relation of interest is whether each dyad was in the same work group or not. In Chapter 2, we created an actor-by-actor (i.e. 75 x 75) binary, symmetric matrix using *Data>Partition to sets* followed by *Data>Affiliations (2-mode to 1-mode)* and saved the dyadic data as "Same_work_groupRows". Notice that the unit of analysis is the dyad, and the dyad is characterized as being in the same group (coded 1) or not (coded 0).

The other relation of interest is whether the members of the dyad are acquainted at the time of the last data collection (wave 4). This matrix is the asymmetric actor-by-actor acquaintanceship data. These data are also measured at the binary level.

We can measure the strength of association between being in the same work group and being acquainted, and test significance, using *Tools>Testing Hypotheses>Dyadic (QAP)>QAP Relational CrossTabs*. The dialog is shown in figure 4.1.

Figure 4.1. UCINET Dialog for Acquaintanceship at End of Course and Same Work Group Crosstab



Relations to Cross-Classify:

C:\Users\apkarian\Dropbox\VTech\Networks Text\Data\Same_work_groupRo
C:\Users\apkarian\Dropbox\VTech\Networks Text\Data\wave4_2011

Parameters

Number of Permutations: 2000

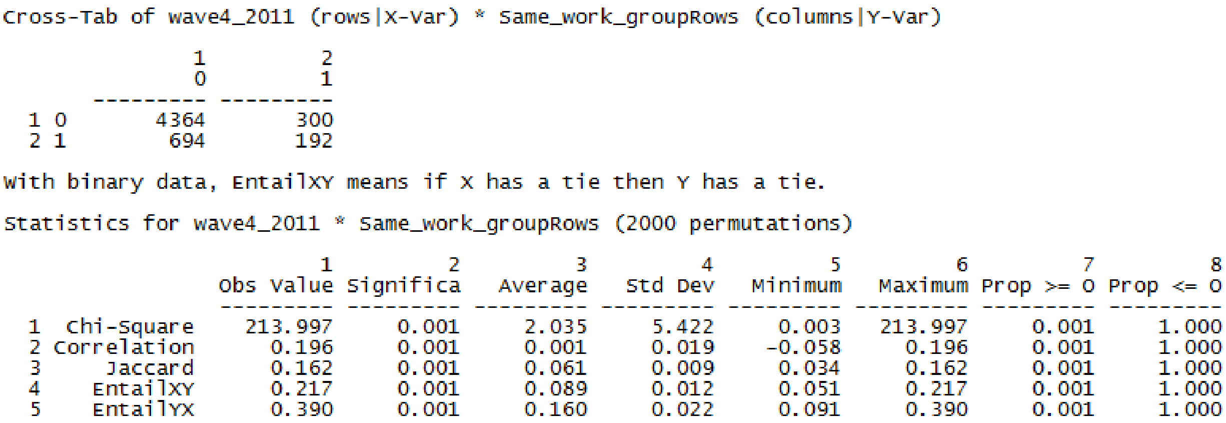
Random Number Seed: 26889

□ Include Missing Values

Output CrossTab:

In the dialog, we used the browsing tool to locate each of the two relational (dyadic) matrices. The number of permutations and random number seed defaults are fine (to recreate our output, use the seed above). We elected to not save the output as a new dataset in this case by leaving “Output CrossTab” empty. Figure 4.2 shows the output.

Figure 4.2. UCINET Output for Acquaintanceship at End of Course and Same Work Group Crosstab



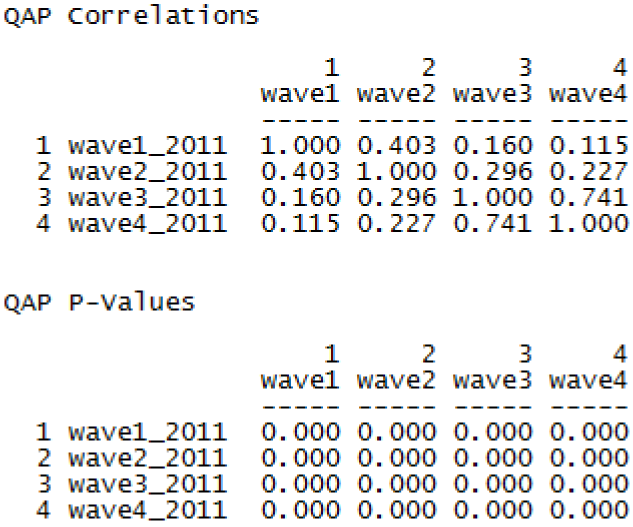
```
Cross-Tab of wave4_2011 (rows|X-Var) * Same_work_groupRows (columns|Y-Var)

                1         2
                0         1
       --------- ---------
  1 0       4364       300
  2 1        694       192

With binary data, EntailXY means if X has a tie then Y has a tie.

Statistics for wave4_2011 * Same_work_groupRows (2000 permutations)

                       1         2         3         4         5         6         7         8
               Obs Value Significa   Average   Std Dev   Minimum   Maximum Prop >= O Prop <= O
               --------- --------- --------- --------- --------- --------- --------- ---------
  1  Chi-Square  213.997     0.001     2.035     5.422     0.003   213.997     0.001     1.000
  2 Correlation    0.196     0.001     0.001     0.019    -0.058     0.196     0.001     1.000
  3     Jaccard    0.162     0.001     0.061     0.009     0.034     0.162     0.001     1.000
  4    EntailXY    0.217     0.001     0.089     0.012     0.051     0.217     0.001     1.000
  5    EntailYX    0.390     0.001     0.160     0.022     0.091     0.390     0.001     1.000
```

The first panel of the output is the relational cross-tabulation itself. It conveys that there were 4,364 dyads that reported no acquaintanceship tie that were also not in the same workgroup. There were 300 dyads among students in the same workgroup that had no claim of acquaintanceship! Of the students who claimed to be acquainted by the final exam, 192 pairs were in the same work group while 694 pairs were not. Simple ratios demonstrate that the likelihood of being acquainted by the final exam is much higher for those in the same work group: 192/(300+192) = 0.39 vs. 694/(4,364+694) = 0.14.

Given this simple joint count, the second panel of the output calculates various measures of association and tests for significant association using permutation tests. The observed Chi-square value in the cross-tab is 213.997. Across the 2,000 cross-tabs generated from randomly permuted data, the average Chi-square statistic was 2.035, with a standard deviation (standard error) of 5.422. Obviously, our observed chi-square is very unlikely to occur in random data. The strength of the association, however, is not terribly impressive (the correlation is 0.20, for example). Clearly, students placed in the same work group were more likely to claim that they were acquainted than those not in the same workgroups, but it doesn't look like the group projects made a really big difference in building even weak-tie networks in the class!

As a second brief example, we can look at the stability of reported acquaintanceship over the four waves using *Tools>Testing Hypotheses>Dyadic (QAP)>QAP Correlation* and choosing all four waves. The resulting Pearson's correlation matrix is shown as figure 4.3.

Figure 4.3. UCINET Generated Correlations of Acquaintanceship Relations over Four Waves



```
QAP Correlations

                    1     2     3     4
                wave1 wave2 wave3 wave4
                ----- ----- ----- -----
  1 wave1_2011  1.000 0.403 0.160 0.115
  2 wave2_2011  0.403 1.000 0.296 0.227
  3 wave3_2011  0.160 0.296 1.000 0.741
  4 wave4_2011  0.115 0.227 0.741 1.000


QAP P-Values

                    1     2     3     4
                wave1 wave2 wave3 wave4
                ----- ----- ----- -----
  1 wave1_2011  0.000 0.000 0.000 0.000
  2 wave2_2011  0.000 0.000 0.000 0.000
  3 wave3_2011  0.000 0.000 0.000 0.000
  4 wave4_2011  0.000 0.000 0.000 0.000
```

As one would expect, the similarity of acquaintanceship structures declines with the length of period of time between measurements. The largest shift in ties appears to occur between waves two and three. Waves one and two are most highly correlated with one another, and the same is true for waves three and four.

4.2.2 Valued Relations

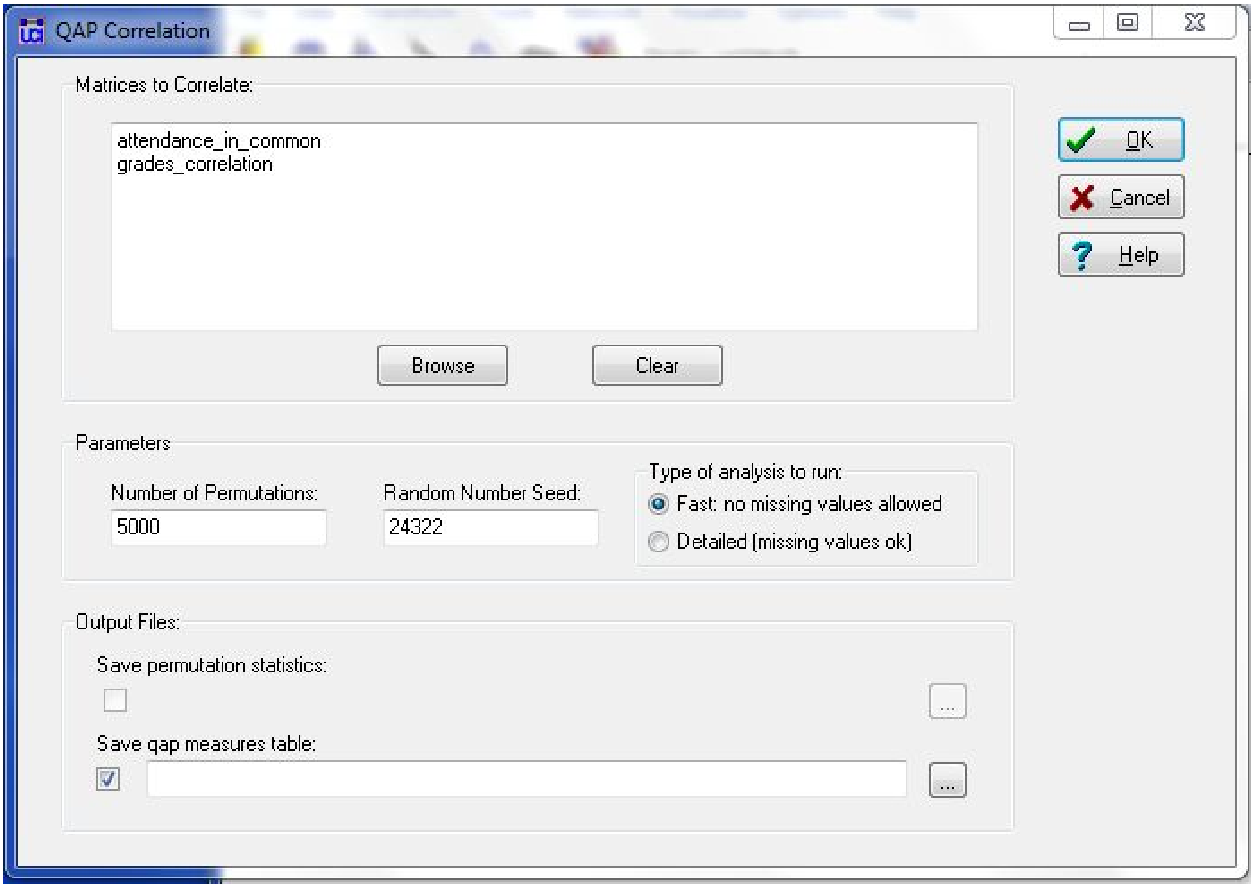
In many cases the relation between the members of a dyad is measured as a quantity, or "valued" relation. Valued relations often indicate the strength of a tie, or some similarity between the two actors (perhaps their nearness or closeness in geographical or network space), or the probability that a tie is present. The natural approach to seeing if two networks of valued relations are associated is to compute the correlation between the tie strengths in one relation with the corresponding tie strengths in the other relation.

Two actors might be expected to be more similar (closer), or be more likely to form social ties if they are frequently co-present in social contexts. That is, actors who have the same pattern of affiliation might be said to have a tie. Using data on which classes each student attended, we used the *Data>Affiliations (2-mode to 1-mode)* tool to create an actor-by-actor matrix of the number of times (potentially 0 to 11) each pair of students had attended the same classes. Dyads with higher values indicate a more similar attendance pattern for the students in the dyad. The same procedure discussed in Chapter 2 (see the dialog in Figure 2.13) can be used to generate this dyadic data.

Let's say we're interested in testing the idea that students who attend the same classes are likely to have similar patterns of exam performance. So, again using the attribute data file, we also can calculate how similar the grades of each pair of students were by using the *Data>Affiliations (2-mode to 1-mode)* tool to create the correlation between students based on the three exam scores (check the box for "Correlation" in the "Affiliations: Convert 2-mode to 1-mode data" dialog). That is, two students are similar, or have a "strong tie" if there is high correlation between scores across the three tests.

Figure 4.4 shows the dialog for calculating the QAP correlation between the similarity in attendance matrix and the correlation of tests matrix. (*Tools>Testing Hypotheses>Dyadic (QAP)>QAP Correlation*). Our research hypothesis is that students who have similar patterns of class attendance are likely to have similar patterns of grades.

Figure 4.4. UCINET Dialog for Correlation between Two Networks



QAP Correlation

Matrices to Correlate:

attendance_in_common
grades_correlation

Parameters

Number of Permutations: 5000

Random Number Seed: 24322

Type of analysis to run:
● Fast: no missing values allowed
○ Detailed (missing values ok)

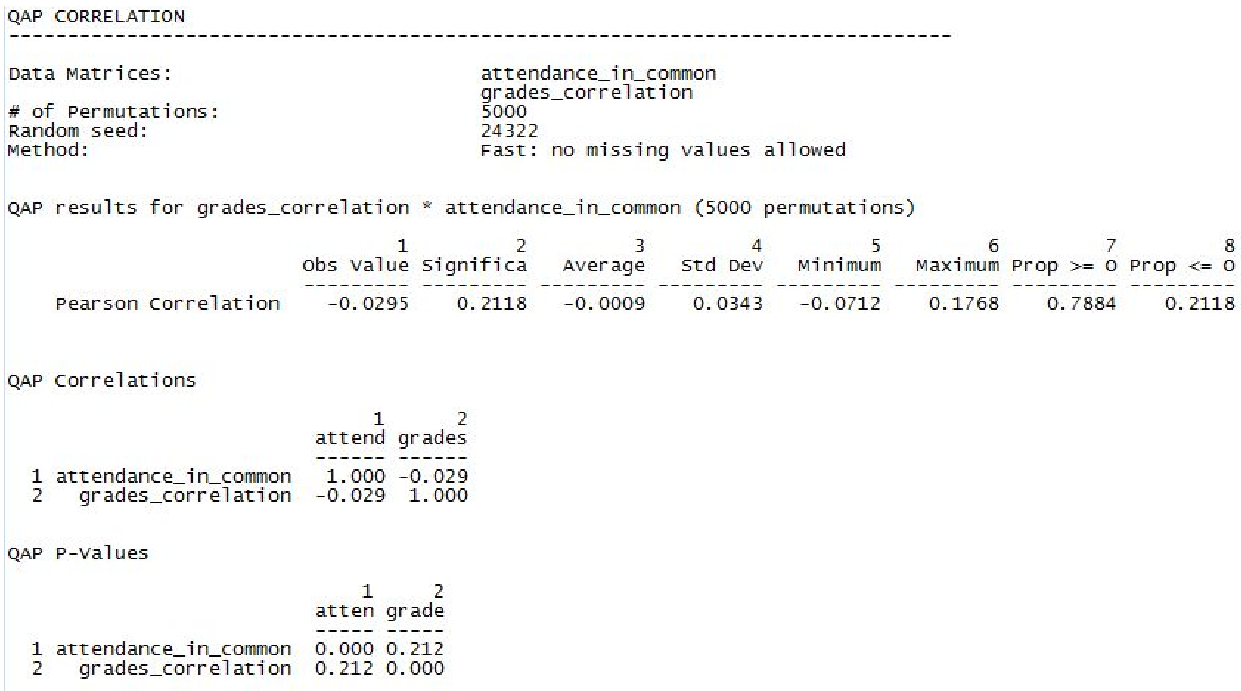
Output Files:

Save permutation statistics: □

Save qap measures table: ☑

The dialog simply asks that the two (or more) dyadic matrices be identified, and allows control over the permutations and saving the results, which are shown in figure 4.5.

Figure 4.5. Dyadic Correlation of Similarities of Class Attendance and Similarity of Exam Grades



```
QAP CORRELATION
--------------------------------------------------------------------------------

Data Matrices:                          attendance_in_common
                                        grades_correlation
# of Permutations:                      5000
Random seed:                            24322
Method:                                 Fast: no missing values allowed

QAP results for grades_correlation * attendance_in_common (5000 permutations)

                                  1          2          3          4          5          6          7          8
                          Obs Value Significa    Average    Std Dev    Minimum    Maximum  Prop >= O  Prop <= O
                          --------- ---------- ---------- ---------- ---------- ---------- ---------- ----------
    Pearson Correlation     -0.0295     0.2118    -0.0009     0.0343    -0.0712     0.1768     0.7884     0.2118


QAP Correlations

                             1      2
                        attend grades
                        ------ ------
  1 attendance_in_common  1.000 -0.029
  2   grades_correlation -0.029  1.000


QAP P-Values

                            1     2
                        atten grade
                        ----- -----
  1 attendance_in_common 0.000 0.212
  2   grades_correlation 0.212 0.000
```

Such a clever hypothesis – and such a disappointing result! We see that there is a non-significant, but negative, correlation between similarity of attendance profiles and similarity of exam grades (*r* = -0.0295, p = 0.2118, two-tail). Across the 5000 random permutations, the average observed correlation was -0.0009, with a standard deviation (standard error) of 0.03243. The amount of overlap that two students had in the classes they attended appears to have nothing to do with achieving similar results on exams. Note, again, that both of these variables are relational or dyadic, in that they describe the relation between two actors and not the attributes of the individual actors.

4.3 Prediction of Networks

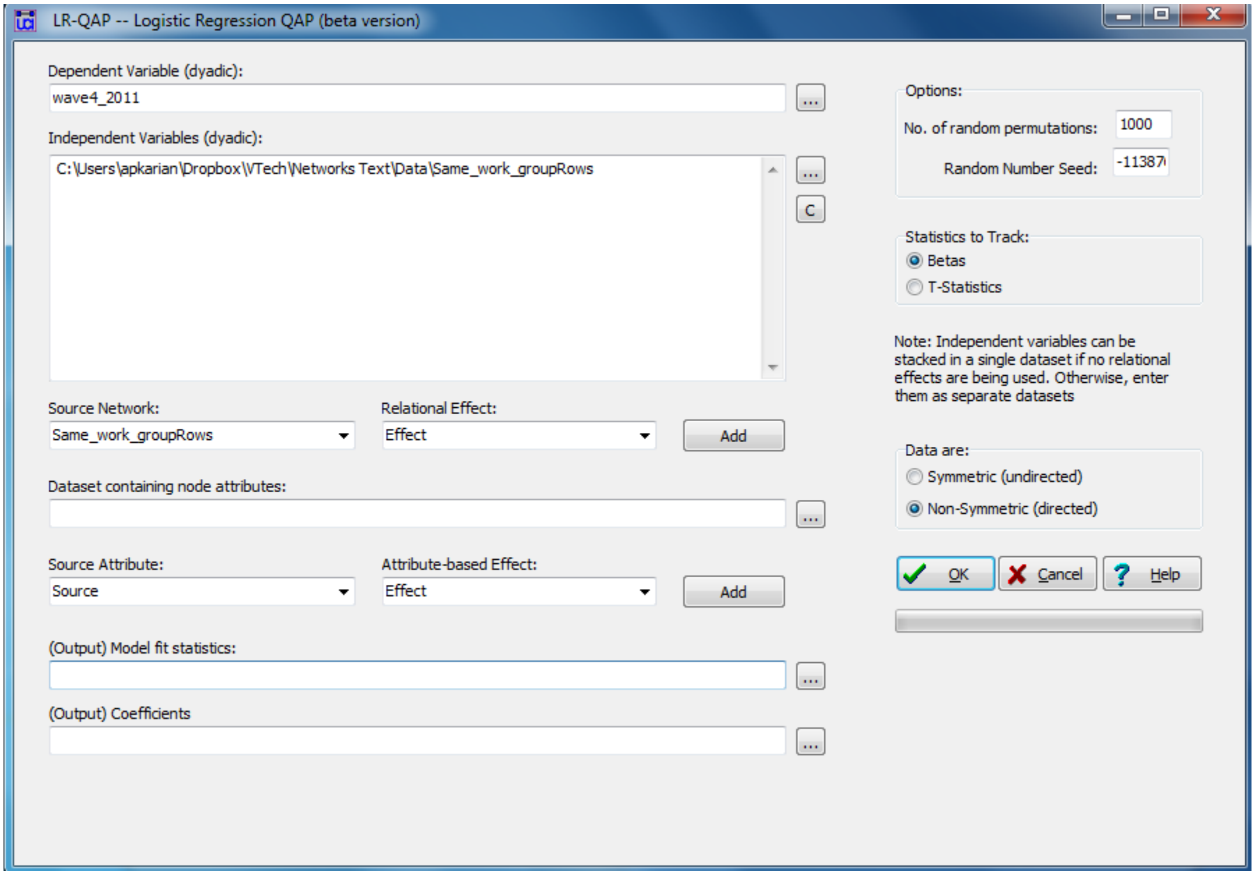
Cross-tabs and correlations are quite adequate for studying many interesting questions about dyadic association. But, if we can treat one relation as dependent and the other(s) as independent, we can apply linear modeling to do prediction and partial association as well.

4.3.1 Binary Relations

If the outcome relation is binary, a natural choice for prediction is binary logistic regression.

Let's re-visit the question of whether being placed in the same work group is associated with being acquainted by the end of the course. The dependent outcome is the binary, asymmetric, matrix of each student nominating others as acquaintances. The independent dyadic variable is being in the same work group, or not. Figure 4.6 shows the dialog of UCINET's *Tools>Testing Hypotheses> Dyadic (QAP)>LR-QAP Logistic Regression (beta).*

Figure 4.6. UCINET Dialog for QAP Logistic Regression of Wave 4 Acquaintanceship by Same Workgroup



LR-QAP -- Logistic Regression QAP (beta version)

Dependent Variable (dyadic): wave4_2011

Independent Variables (dyadic): C:\Users\apkarian\Dropbox\VTech\Networks Text\Data\Same_work_groupRows

Source Network: Same_work_groupRows

Relational Effect: Effect

Dataset containing node attributes:

Source Attribute: Source

Attribute-based Effect: Effect

(Output) Model fit statistics:

(Output) Coefficients

Options:

No. of random permutations: 1000

Random Number Seed: -11387

Statistics to Track:

● Betas

○ T-Statistics

Note: Independent variables can be stacked in a single dataset if no relational effects are being used. Otherwise, enter them as separate datasets

Data are:

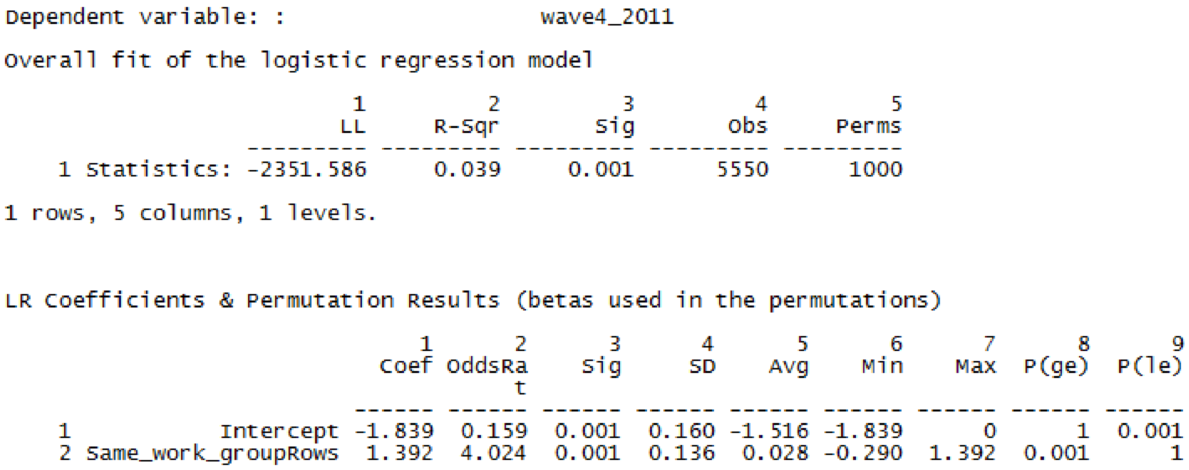
○ Symmetric (undirected)

● Non-Symmetric (directed)

This is a very complicated looking dialog for such a simple question! The reason is that there is much more in this tool than we will be using right now. A bit later on, we will look at a more complex approach to modeling and prediction of networks – exponential random graph (ERG) modeling. The dependent relation or network is our non-symmetric (see the selection of symmetric/non-symmetric on the lower right in the dialog) wave 4 acquaintanceships. Our single "independent" network or dyadic variable is the matrix of "in the same work group." At present, we won't use the "relational effect" or "attribute-based effect" parts of this tool. Note that the regression output can be saved with the output file boxes at the bottom of the dialog.

Figure 4.7 shows the output of our logistic regression model. By the way, ERG models with permutation can take a rather long time to run. So, be (reasonably) patient.

Figure 4.7. UCINET Output for QAP Logistic Regression of Wave 4 Acquaintanceship by Same Work Group



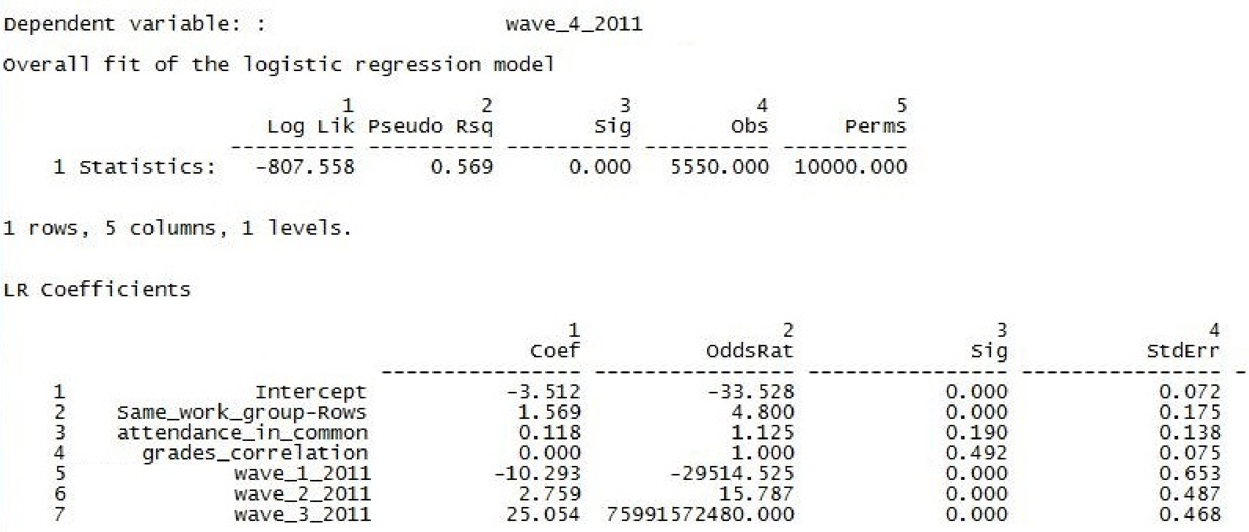
```
Dependent variable: :                   wave4_2011

Overall fit of the logistic regression model

                              1          2          3          4          5
                             LL      R-Sqr        Sig        Obs      Perms
                     ---------- ---------- ---------- ---------- ----------
    1 Statistics:  -2351.586      0.039      0.001       5550       1000

1 rows, 5 columns, 1 levels.


LR Coefficients & Permutation Results (betas used in the permutations)

                                   1      2      3      4      5      6      7      8      9
                                Coef OddsRa    Sig     SD    Avg    Min    Max  P(ge)  P(le)
                                          t
                              ------ ------ ------ ------ ------ ------ ------ ------ ------
    1             Intercept   -1.839  0.159  0.001  0.160 -1.516 -1.839      0      1  0.001
    2 Same_work_groupRows      1.392  4.024  0.001  0.136  0.028 -0.290  1.392  0.001      1
```

The output first reports the overall goodness-of-fit of the model. The log-likelihood is given, along with the likelihood-based pseudo-R square statistic. Note that the N is 5550, or the number of directed pairs formed by 75 nodes. We can conclude that being in the same work group does affect the log-odds of naming another member as an acquaintance (the coefficient is significant; p = 0.001), but that this tendency explains a tiny proportion of the variation in likelihood of being acquainted (pseudo-R square = 0.039).

The regression coefficients show the additive effects on log odds (Coef), or multiplicative effects on odds (OddsRat) of naming alter as an acquaintance. The simple summary here is that being in the same work group as alter multiplies the odds that ego will name them as an acquaintance by about 4 times (4.024). The average regression coefficient in 10,000 random permutations of the data was 0.028, and the standard error was 0.136.

Since we are now working in the generalized linear modeling framework, there is no difficulty in adding additional variables as partial explanations or control variables. In figure 4.8, we've added whether an acquaintanceship was reported at wave 1, wave 2, or wave 3, as well as having similar attendance records and similar exam grades as additional predictors of acquaintanceship at the end of the course.

Figure 4.8. UCINET Output for QAP Logistic Regression of Wave 4 Acquaintanceship on Multiple Predictors



```
Dependent variable: :                  wave_4_2011
Overall fit of the logistic regression model

                       1          2          3          4          5
                 Log Lik Pseudo Rsq        Sig        Obs      Perms
              ---------- ---------- ---------- ---------- ----------
    1 Statistics:   -807.558      0.569      0.000   5550.000  10000.000

1 rows, 5 columns, 1 levels.

LR Coefficients

                                            1                2                3                4
                                         Coef          OddsRat              Sig           StdErr
                           ---------------- ---------------- ---------------- ---------------- -
    1              Intercept           -3.512          -33.528            0.000            0.072
    2    Same_work_group-Rows           1.569            4.800            0.000            0.175
    3    attendance_in_common           0.118            1.125            0.190            0.138
    4      grades_correlation           0.000            1.000            0.492            0.075
    5             wave_1_2011         -10.293       -29514.525            0.000            0.653
    6             wave_2_2011           2.759           15.787            0.000            0.487
    7             wave_3_2011          25.054   75991572480.000            0.000            0.468
```

The overall fit of the model is now very much better. However, most of the improvement is due to including the presence of ties earlier in the term as predictors of ties at the end of the term. Note that, while significant, the autoregression of current social ties on earlier ones has substantial collinearity and produces some very unstable results (UCINET's routines do not include VIF or other collinearity statistics). Our "test" variables can be interpreted as attempting to predict the presence or absence of ties that would not have been expected based on the trend in the development of ego's network. Having similar grades has no discernible partial effects at all. Having a similar attendance profile is very slightly (but not significantly) partially associated with being acquainted. But, the positive effect of being in the same work group is strengthened somewhat by controlling for the other variables.

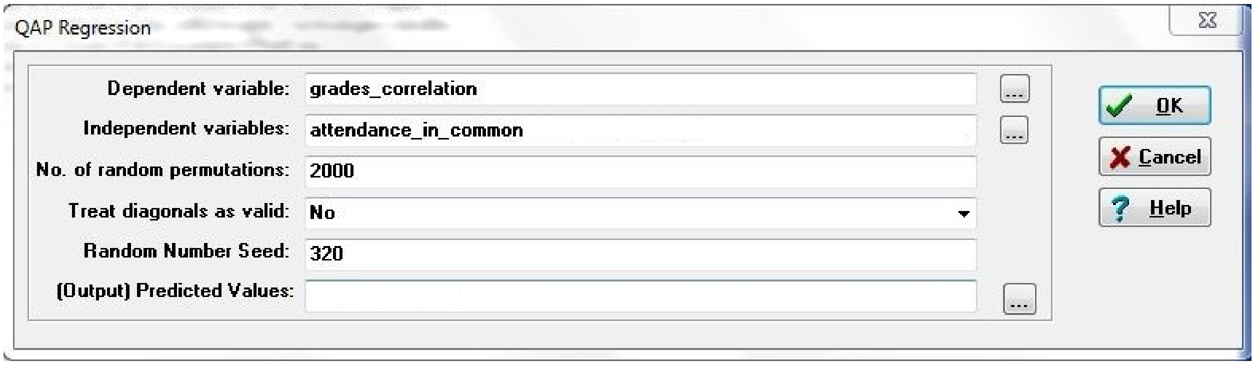
The UCINET tool for predicting a binary dyadic dependent variable is quite easy to use for simple models. The prediction of a binary dyadic relation from any combination of independent variables can also be approached with specialized software that takes graph structure into account (exponential random graph models), or multi-level generalized linear models. We'll return to these more general tools in later chapters. In the example above (Figure 4.8), earlier observations were used as predictors of later relations; the data are actually panel data. Specialized models for studying change in dyadic relational variables are also available (e.g. Siena).

4.3.2 Valued Relations

The same logic and approaches can be applied where the outcome dyadic variable is an interval-ratio level measure of dyadic tie-strength. Figure 4.9 shows the dialog of the UCINET dyadic regression tool (*Tools>Testing Hypotheses>Dyadic (QAP)>MR QAP Linear Regression>Original (Y Permutation) method*).

In this example, we are attempting to predict the correlation between the exam grades of the two members of each dyad based on the similarity of their attendance patterns.

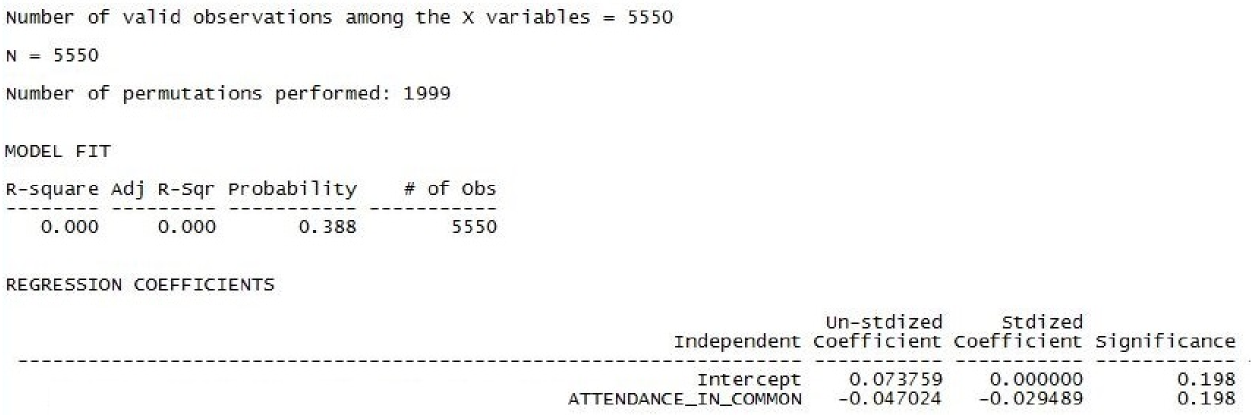
Figure 4.9. UCINET Dialog for QAP Regression Predicting Grade Similarity by Common Attendance



Dependent variable: grades_correlation
Independent variables: attendance_in_common
No. of random permutations: 2000
Treat diagonals as valid: No
Random Number Seed: 320
(Output) Predicted Values:

The regression dialog is simpler, and has a slightly different appearance. The dependent network is selected in the first box. One or more independent dyadic variables (networks) are selected in the second. Figure 4.10 shows the resulting output.

Figure 4.10. UCINET Output of Grade Similarity by Common Attendance

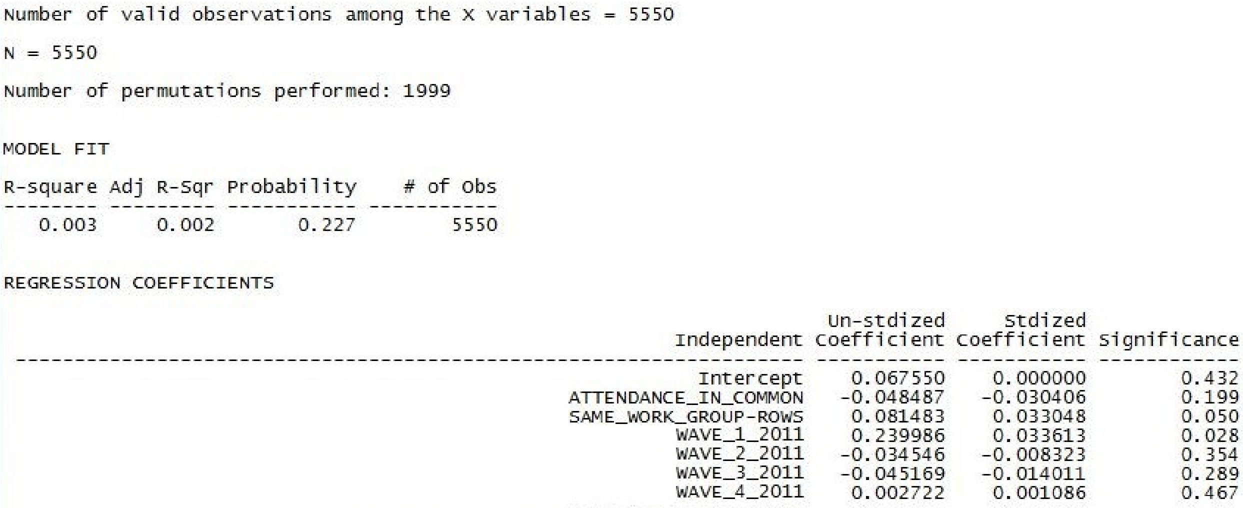


```
Number of valid observations among the X variables = 5550

N = 5550

Number of permutations performed: 1999

MODEL FIT

R-square Adj R-Sqr Probability    # of Obs
-------- --------- ----------- -----------
   0.000     0.000       0.388        5550

REGRESSION COEFFICIENTS

                                                               Un-stdized      Stdized
                                                 Independent Coefficient Coefficient Significance
 ----------------------------------------------------------------- ----------- ----------- ------------
                                                   Intercept    0.073759    0.000000        0.198
                                          ATTENDANCE_IN_COMMON   -0.047024   -0.029489        0.198
```

The results are consistent with those found in Figure 4.5. The model fit panel tells us that any association between similar grades and similar attendance could easily be explained by random processes (p = 0.388). We really should not bother to look at the regression slope. If we did, we would note that the tendency in the data (a standardized slope of -0.03) contradicts our research hypothesis.

However, there may be other factors suppressing the true relationship between attendance similarity and grade correlation so let's add some control variables. Figure 4.11 shows the multiple regression output of predicting similarity of grades from common attendance patterns, while controlling for being in the same work group and the presence of acquaintanceship at any point during the academic term.

Figure 4.11. UCINET Output of Grade Similarity by Multiple Predictors



```
Number of valid observations among the X variables = 5550

N = 5550

Number of permutations performed: 1999

MODEL FIT

R-square Adj R-Sqr Probability    # of Obs
-------- --------- ----------- -----------
   0.003     0.002       0.227        5550

REGRESSION COEFFICIENTS

                                                                    Un-stdized      Stdized
                                                       Independent Coefficient Coefficient Significance
  ----------------------------------------------------------------- ----------- ----------- ------------
                                                         Intercept    0.067550    0.000000        0.432
                                               ATTENDANCE_IN_COMMON   -0.048487   -0.030406        0.199
                                               SAME_WORK_GROUP-ROWS    0.081483    0.033048        0.050
                                                        WAVE_1_2011    0.239986    0.033613        0.028
                                                        WAVE_2_2011   -0.034546   -0.008323        0.354
                                                        WAVE_3_2011   -0.045169   -0.014011        0.289
                                                        WAVE_4_2011    0.002722    0.001086        0.467
```

The results are not impressive. Looking at the overall goodness of fit, we should conclude that the similarity of the grades between ego and alter is essentially random with respect to these predictors. Poking into the partial slopes (which we really shouldn't do), we see that direct acquaintanceship ties only seem to matter at the beginning of the quarter – probably reflecting friendship relationships that existed before starting the class. Being in the same work-group also appears to be associated with having more similar grades. Just as in the bivariate model, attending the same classes is not associated with getting similar grades.

4.4 Summary

In this short chapter we've considered the question of how to study whether the pattern of one set of ties among a given set of actors is similar to the pattern of another set of ties among the same actors. That is, the association of two (or more) networks.

The approach that we've looked at here examines each network as a collection of the relations of all possible pairs (dyads) in the network. This treats the dyad as the unit of observation, and the relation between them as the variable to be studied. Approached this way, the conventional tools of tables, correlation, and regression can all be applied to the association between networks.

The inferential statistical question in examining the relationships among two or more networks is not really one of generalization to a population. Rather, it is whether the observed degree of correspondence or similarity between two relations among the actors in a network could have happened by random processes. So, for testing hypotheses about the relations among networks, the permutation method for estimating standard errors is ideal.

The tools that we've looked at in this chapter are actually quite simple. But, they can be useful for many interesting questions. Later on, we'll see more complex approaches (exponential random graphs models and multi-level models) to examining entire networks as dependent variables that allow for much more complex and interesting hypotheses.

We've now looked at how to examine the association between two or more attributes when they are observed on actors embedded in a network (Chapter 3). We've also looked at how to examine the association among two or more networks (this chapter).

Logically, this raises the question of how one might study the association between attributes and networks – i.e. the association between nodal and dyadic relational variables. Read on!

Chapter 5. Association Between Attributes and Networks

In the previous two chapters we looked at studying association between two or more attributes of actors in a network (Chapter 3), and at studying association between two or more networks (Chapter 4). In this chapter, we'll examine how to look at relationships between an attribute and a network.

5.1 Attributes and Networks

What does it mean to examine the relationship between an attribute (node-level) variable, and a relational (dyadic-level) variable? This sounds a bit abstract. Let's consider a couple of examples.

Citing the "homophily" principle (birds of a feather flock together), we might suppose that the women students from the classroom data (see Chapters 2-4) would be more likely to be acquainted with other women students than with men students in the class, and that men students were more likely to be acquainted with other men students. A student's gender is a nodal, or attribute, or monadic variable. Whether acquaintanceships are woman-man, woman-woman, or man-man is a relational level variable. The variation being described is between pairs of nodes, not individuals. So, hypothesizing that there is gender homophily in relationships is actually making a prediction about the relationship between a nodal variable and a dyadic variable.

Citing the "social learning" principle (people are likely to be influenced by, and learn from those with whom they have social ties), we might suppose that students who are acquainted may have more similar grades. It may be that the students practice "network selection" in forming and breaking ties to increase similarity. Alternatively, it may be that students with social ties influence one another to become more similar. In either case, the correlation between students' individual attributes (grades) is hypothesized to be a function of how far apart they are in the acquaintanceship network (which is a dyadic variable describing each pair of students).

In SNA, hypotheses that link the individual level and network level of analysis are common. Individual social actors may "select" which ties to make or break based on their own (or others) attributes. When social ties exist, individuals are influenced by the attributes of the others to whom they have ties. In SNA, nodal attributes may be independent variables that determine how an actor becomes embedded in the network (which dyads they are a part of). Nodal attributes may also be dependent variables that are affected by how the actor is embedded.

In later chapters, we will take a look at modeling these kinds of influence and selection processes that connect individuals and their networks. Often, our questions about the relation between an attribute and a network are quite simple and can be addressed with some rather simple tools. We'll look first at a variety of approaches that are helpful when the attribute in question is categorical. For example, gender might be an independent attribute variable that affects the likelihood of acquaintanceship, or acquaintanceship might be an independent variable that affects test performance. Then, we'll look at a rather different approach – borrowed from geo-statistics—that is helpful when the attribute being studied is interval-ratio.

5.2 Categorical Attributes

Some of the most obvious, and also most important, questions about attributes and networks ask whether pre-existing nodal attributes determine how individuals are embedded in a network. Of course, the same kind of questions can be asked in reverse: to what extent does the way that an individual is embedded affect their individual attributes or behaviors.

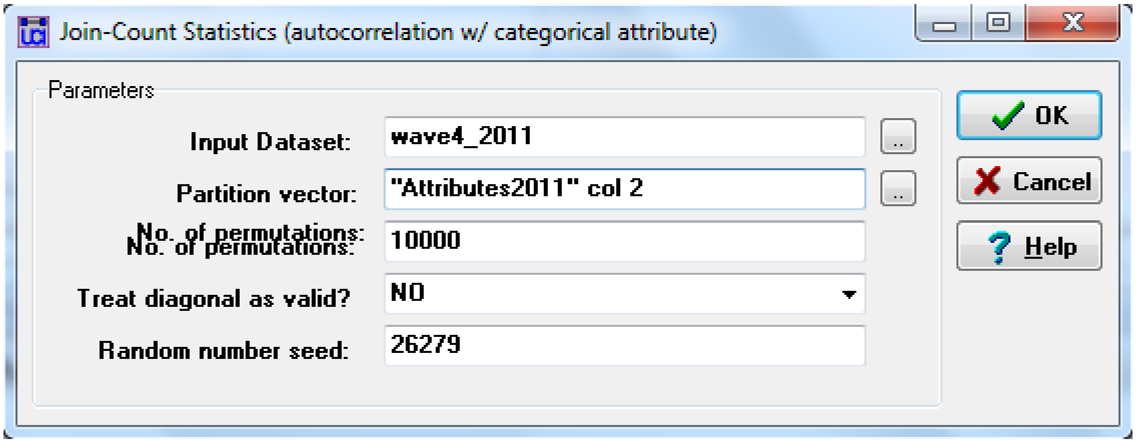
There are several approaches that one can take. If the nodal attribute has two values (e.g. man or woman, passed the test or didn't), a group-comparison approach can be used. If the nodal attribute has multiple values (e.g. ethnicity, which work group a person was in, which letter grade they earned on the final), a cross-tab approach can be applied. When the dyadic variable is either categorical or continuous, ANOVA density models or "block" models are a very interesting and useful approach.

5.2.1 Two Groups: Joint Counts

Let's suppose that our nodal variable is binary, and that our dyadic variable is also binary. We'll consider two examples.

First, are there differences between men and women students in the gender of the others that they are acquainted with? Here, our nodal independent variable is whether a person identifies as a man or a woman. Our dyadic dependent variable is the count of the number of dyads that the individual is involved with that have a tie to a woman or a man. Here's how we can test this hypothesis using UCIENET's *Tools>Testing Hypotheses>Mixed Dyadic/Nodal>Categorical Attribute>Joint Count*. The dialog box used to set up the test is shown in figure 5.1.

Figure 5.1. UCINET Dialog for Testing Acquaintanceship Patterns by Gender



Join-Count Statistics (autocorrelation w/ categorical attribute)

Parameters

Input Dataset: wave4_2011

Partition vector: "Attributes2011" col 2

No. of permutations: 10000

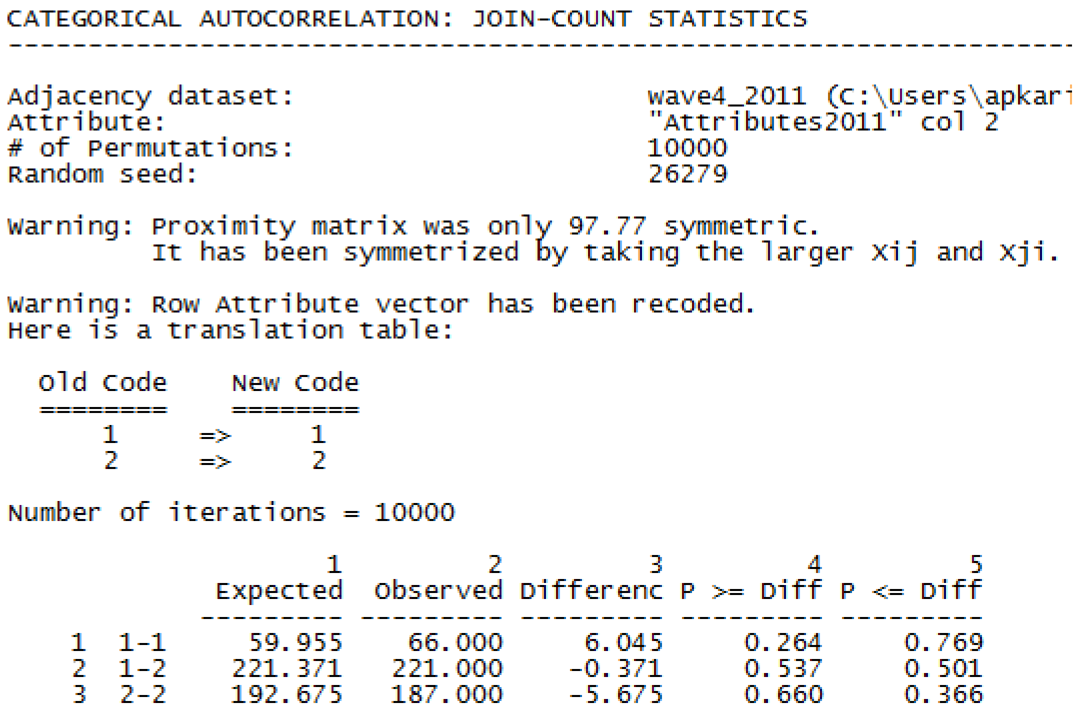
Treat diagonal as valid? NO

Random number seed: 26279

This looks a little mysterious. The "input dataset" is the location of the relational or dyadic variable of interest in the problem. Here, we have selected the matrix that describes whether each individual does, or does not claim to be acquainted with each other individual at the end of the course (wave 4). The "partition vector" is the column of the attribute data file that describes whether each individual is a man (coded 1) or a woman (coded 2). This happens to be the second column of our attributes data file.

To test whether there are differences between men and women in the likelihood that they form dyads with men or women, we will be using permutation tests, so the remaining parts of the dialog are defaults. Now, consider the output, shown in figure 5.2.

Figure 5.2. UCINET Output for Testing Acquaintanceship Patterns by Gender



```
CATEGORICAL AUTOCORRELATION: JOIN-COUNT STATISTICS
------------------------------------------------------------------

Adjacency dataset:                        wave4_2011 (C:\Users\apkari
Attribute:                                "Attributes2011" col 2
# of Permutations:                        10000
Random seed:                              26279

Warning: Proximity matrix was only 97.77 symmetric.
         It has been symmetrized by taking the larger Xij and Xji.

Warning: Row Attribute vector has been recoded.
Here is a translation table:

  Old Code     New Code
  ========     ========
      1     =>     1
      2     =>     2

Number of iterations = 10000

                         1          2          3          4          5
                  Expected   Observed Differenc P >= Diff P <= Diff
                 ---------- ---------- ---------- ---------- ----------
    1  1-1          59.955     66.000      6.045      0.264      0.769
    2  1-2         221.371    221.000     -0.371      0.537      0.501
    3  2-2         192.675    187.000     -5.675      0.660      0.366
```

The procedure forms a cross-tabulation (not displayed). On one axis of the table is the attribute variable of whether the node is a man or a woman. On the other axis of the table is whether the other member of the dyad is a man or a woman. The output table "Expected" column shows the "joint count" of the number of man-man (1-1), man-woman (1-2) and woman-woman (2-2) dyads that we would expect to see if each respondent's ties were distributed at random across all other persons (i.e. the number of each type of dyad we'd expect if the likelihood of a tie to a man or a woman was independent of ego's gender). The "Observed" column shows how many such dyads were actually observed in the data.

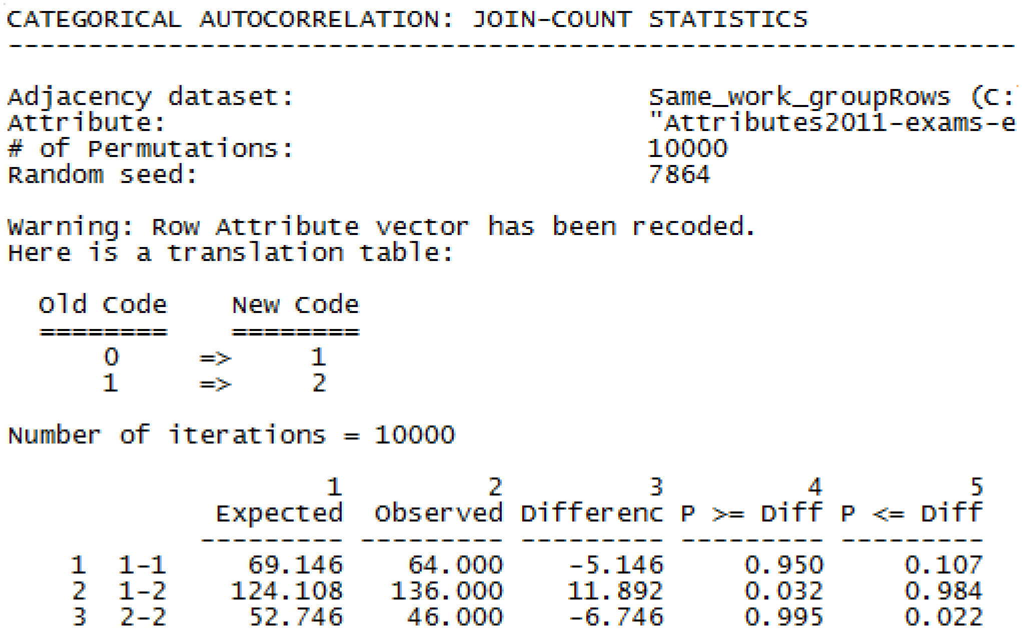
In this example, we see that men students are more likely to affiliate with other men students than we would have expected under independence (they form 66 dyads with other men, rather than the 60 we would have expected if gender was irrelevant). Women students display a tendency to fewer ties with other women students than we would have expected (187 observed, versus 193 expected). However, differences this large in expected and observed counts happen rather frequently in joint-counts formed from randomly permuted data (that is, randomly assigning cases to dyadic ties with men or women). The p-level for man-man departure from independence is 0.264. For woman-woman dyads, the p-level is 0.366.

In a second example, let's treat the individual attribute as dependent. First, let's divide individuals into a partition (attribute) that measures whether they did, or did not, achieve a score of 70% or better on the final exam. This is an opportunity to demonstrate the power of UCINET's Excel Matrix Editor. First open the editor by choosing *Data>Data editors>Excel Matrix Editor* or by clicking on the Excel icon in UCINET's button bar across the top of the main window. Then choose *Open>Open UCINET dataset* and select the "Attributes_2011" file. In cell M1, create a new variable called "PassE3." This variable will be coded "2" if the student received a score of 70 or better on E3 and "1" otherwise (UCINET's *Joint* command sometimes has problems with "0"s for dichotomous attributes, so it's best to avoid them). In cell M2, type "=if(J2>=70,2,1)" and hit Enter. Cell M2 now looks at cell J2 (which is student AD's exam 3 score), checks whether it's greater or equal to 70 (which it is) and codes either 2 or 1 for pass or fail respectively (it should read "2"). Now select M2 again, grab the small black square in the lower right corner and drag the cursor all the way down until you stop on cell M76. This now follows the same procedure for every student in the class and our new variable, PassE3, is ready to be used. Now save the data as a UCINET dataset and call it "Attributes_2011-PassE3," to be used shortly.

For our independent variable, we'll use a dyadic matrix that shows whether each pair of students were (coded "1") or were not (coded "0") in the same workgroup. We already did this in Chapter 2, where we created an actor-by-actor (i.e. 75 x 75) binary, symmetric matrix using *Data>Partition to sets* followed by *Data>Affiliations (2-mode to 1-mode)* and saved the dyadic data as "Same_work_groupRows". Our research hypothesis is that students who are in the same workgroup are more likely to achieve similar outcomes on the final exam.

The dialog is the same as in figure 5.1, with “Same_work_groupRows” specified as the input dataset (dyadic variable), and “"Attributes2011-PassE3" col 12” (pass/fail on the exam) as the partition vector (nodal variable). The output is shown in figure 5.3.

Figure 5.3. UCINET Output Testing if Being in the Same Work Group Affects Passing the Final Exam



```
CATEGORICAL AUTOCORRELATION: JOIN-COUNT STATISTICS
--------------------------------------------------------------

Adjacency dataset:                     Same_work_groupRows (C:
Attribute:                             "Attributes2011-exams-e
# of Permutations:                     10000
Random seed:                           7864

Warning: Row Attribute vector has been recoded.
Here is a translation table:

  Old Code    New Code
  ========    ========
      0    =>     1
      1    =>     2

Number of iterations = 10000

                    1         2         3         4         5
             Expected  Observed Differenc P >= Diff P <= Diff
            --------- --------- --------- --------- ---------
    1  1-1     69.146    64.000    -5.146     0.950     0.107
    2  1-2    124.108   136.000    11.892     0.032     0.984
    3  2-2     52.746    46.000    -6.746     0.995     0.022
```

The “1-1” row shows dyads where both members failed the exam. There were 64 such dyads, slightly fewer than the number we would have expected under the null hypothesis of independence. Evidence against our theory. The “2-2” row is the count of dyads where both students passed the exam. There were significantly fewer (46) dyads than we would have expected under independence (53). And, the row “1-2” shows that there were 136 pairs where one member passed and the other failed. This happened more frequently than we would have expected under independence. It looks like we were wrong. If anything, there is a tendency for members of the same work groups to have different, rather than similar exam outcomes!

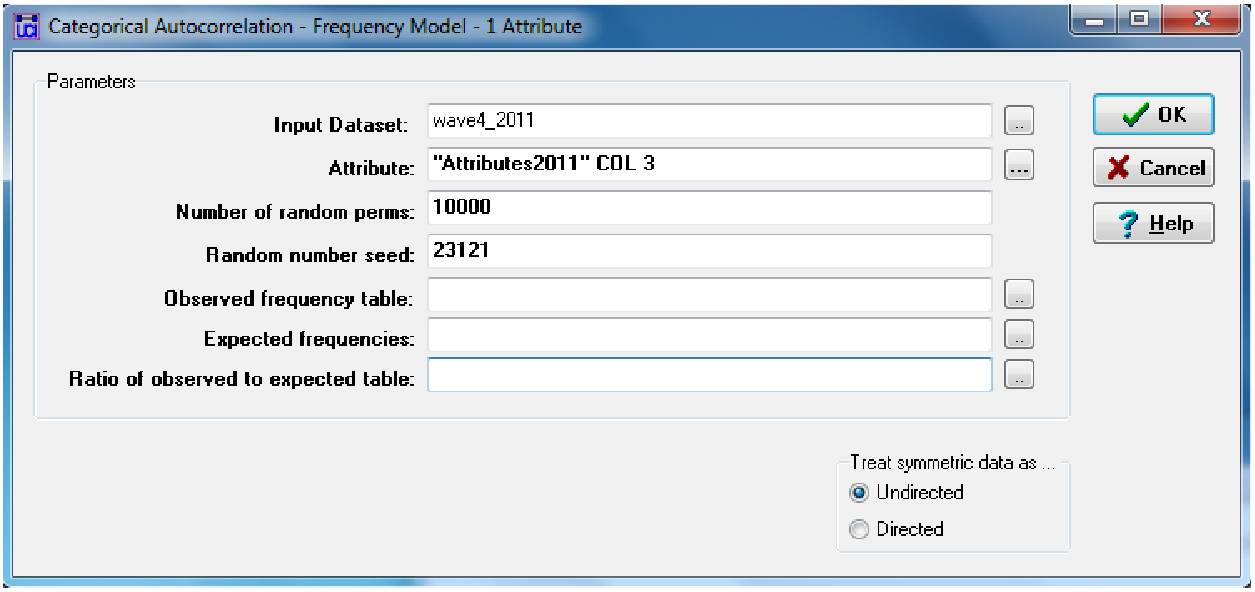
To reprise: in these examples we are examining whether there is an association between individual attributes and the frequency of involvement in dyads of particular kinds. In the first example, the individual's attribute was the independent variable, gender, and the dyadic attribute was the gender mix of the dyads that they were embedded in. In the second example, the individual attribute of interest was dependent: whether the individuals involved in the dyad passed or did not pass the exam. The independent variable was the dyadic variable of being in the same work group or not.

5.2.2 Multiple Groups: A Contingency Table Approach

We can expand the same kind of thinking to cases where the partition or attribute variable contains multiple categories. That is, we can look at the association between a dyadic or relational variable and a multi-valued attribute or nodal variable.

Let's ask whether there is an association between which workgroups students were assigned to, and whether there are ties between them. Students were assigned to one of ten groups (a nodal variable), and were acquainted or not with each other student by the end of the term (a dyadic variable). We hypothesize that students are likely to have a higher density or probability of forming dyads with other members of their work group. Figure 5.4 shows the dialog for the UCINET "relational contingency table" tool (*Tools>Testing Hypotheses>Mixed Dyadic/Nodal>Categorical Attribute>Relational Contingency Tables*).

Figure 5.4. UCINET Dialog for Testing Acquaintanceship Differences among Work-Groups



Categorical Autocorrelation - Frequency Model - 1 Attribute

Parameters

Input Dataset: wave4_2011

Attribute: "Attributes2011" COL 3

Number of random perms: 10000

Random number seed: 23121

Observed frequency table:

Expected frequencies:

Ratio of observed to expected table:

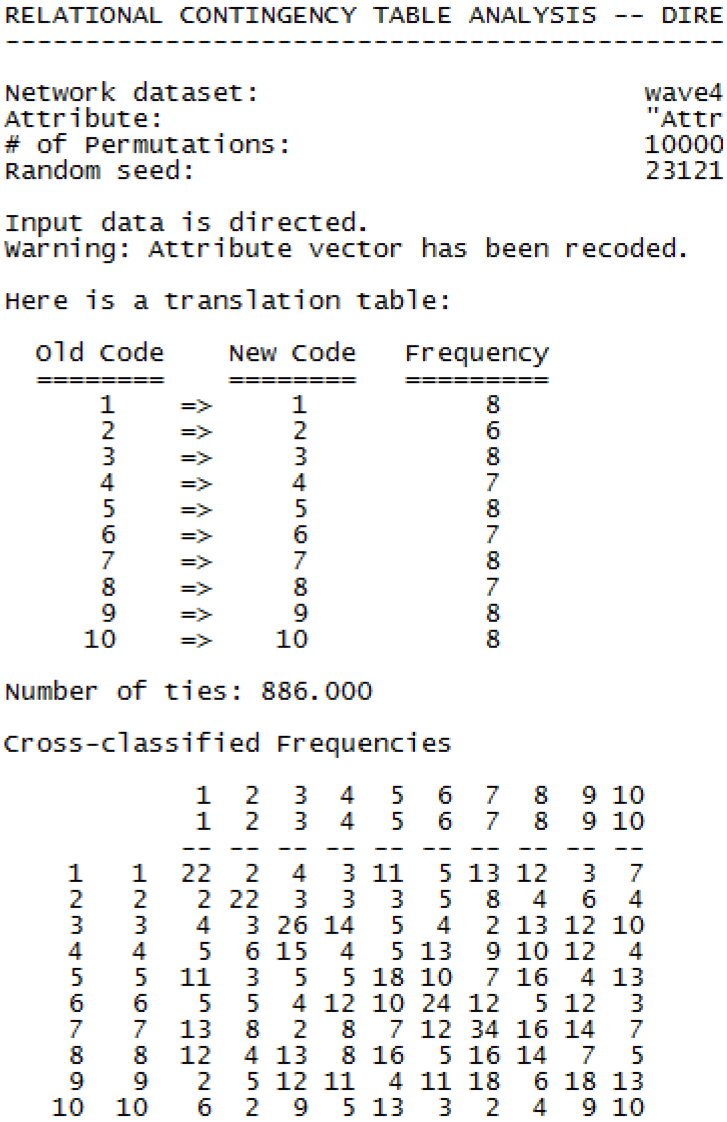
Treat symmetric data as ...

● Undirected

○ Directed

The "input dataset" is the square student-by-student dyadic acquaintanceship relation (we use wave 4, taken during the final exam). The "attribute" or categorical nodal variable is column 3 from the 2011 "Attributes" dataset, which contains the number of the work group to which each student was assigned. Again, there are controls for the permutation trials, and the option to save parts of the output as datasets for further processing. Figures 5.5.1, 5.5.2, and 5.5.3 show the full (somewhat lengthy) output.

Figure 5.5.1. UCINET Output for Testing Acquaintanceship Differences among Work-Groups (Part 1)

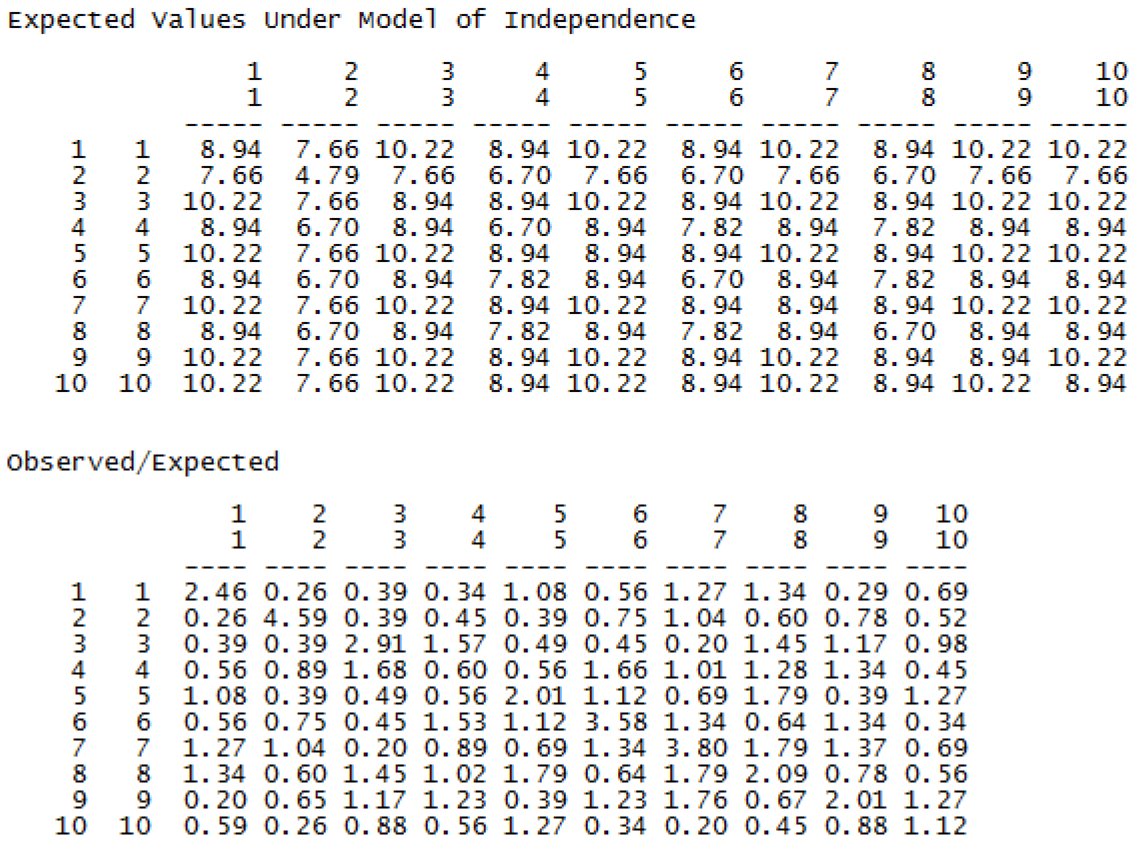


```
RELATIONAL CONTINGENCY TABLE ANALYSIS -- DIRE
---------------------------------------------

Network dataset:                        wave4
Attribute:                              "Attr
# of Permutations:                      10000
Random seed:                            23121

Input data is directed.
Warning: Attribute vector has been recoded.

Here is a translation table:

  Old Code    New Code   Frequency
  ========    ========   =========
      1    =>     1          8
      2    =>     2          6
      3    =>     3          8
      4    =>     4          7
      5    =>     5          8
      6    =>     6          7
      7    =>     7          8
      8    =>     8          7
      9    =>     9          8
     10    =>    10          8

Number of ties: 886.000

Cross-classified Frequencies

              1  2  3  4  5  6  7  8  9 10
              1  2  3  4  5  6  7  8  9 10
             -- -- -- -- -- -- -- -- -- --
    1   1    22  2  4  3 11  5 13 12  3  7
    2   2     2 22  3  3  3  5  8  4  6  4
    3   3     4  3 26 14  5  4  2 13 12 10
    4   4     5  6 15  4  5 13  9 10 12  4
    5   5    11  3  5  5 18 10  7 16  4 13
    6   6     5  5  4 12 10 24 12  5 12  3
    7   7    13  8  2  8  7 12 34 16 14  7
    8   8    12  4 13  8 16  5 16 14  7  5
    9   9     2  5 12 11  4 11 18  6 18 13
   10  10     6  2  9  5 13  3  2  4  9 10
```

First we see the frequencies of the partition (work group number) variable. The groups varied in size between 6 and 8 students each. Next, we see the raw frequencies of the dyads formed by students in each group (row) with students in their own and other groups (column). For example, the students in group 1 claimed acquaintances 22 times with other members of their own group, only twice with a student from workgroup 2, eleven times with students in workgroup 5, and so on.

Note that the main diagonal looks pretty dense. That is, it looks like there is a tendency for students in each group to form dyads with other students in their same work-group. Also note that this tendency toward "homophily" isn't the same across all work groups (note the diagonal value for group 4).

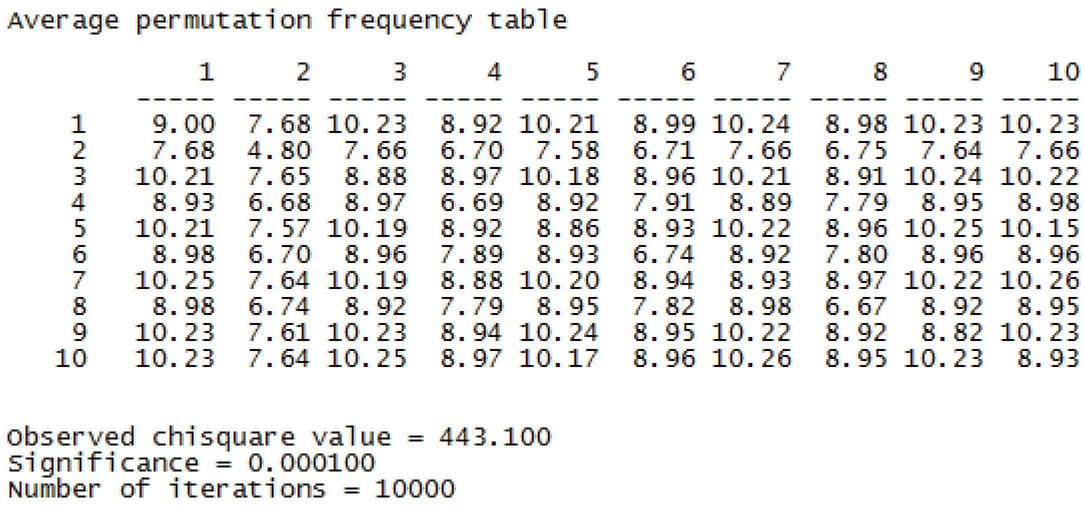
Figure 5.5.2. UCINET Output for Testing Acquaintanceship Differences among Work-Groups (Part 2)



```
Expected Values Under Model of Independence

                 1     2     3     4     5     6     7     8     9    10
                 1     2     3     4     5     6     7     8     9    10
             ----- ----- ----- ----- ----- ----- ----- ----- ----- -----
  1   1       8.94  7.66 10.22  8.94 10.22  8.94 10.22  8.94 10.22 10.22
  2   2       7.66  4.79  7.66  6.70  7.66  6.70  7.66  6.70  7.66  7.66
  3   3      10.22  7.66  8.94  8.94 10.22  8.94 10.22  8.94 10.22 10.22
  4   4       8.94  6.70  8.94  6.70  8.94  7.82  8.94  7.82  8.94  8.94
  5   5      10.22  7.66 10.22  8.94  8.94  8.94 10.22  8.94 10.22 10.22
  6   6       8.94  6.70  8.94  7.82  8.94  6.70  8.94  7.82  8.94  8.94
  7   7      10.22  7.66 10.22  8.94 10.22  8.94  8.94  8.94 10.22 10.22
  8   8       8.94  6.70  8.94  7.82  8.94  7.82  8.94  6.70  8.94  8.94
  9   9      10.22  7.66 10.22  8.94 10.22  8.94 10.22  8.94  8.94 10.22
 10  10      10.22  7.66 10.22  8.94 10.22  8.94 10.22  8.94 10.22  8.94


Observed/Expected

                1    2    3    4    5    6    7    8    9   10
                1    2    3    4    5    6    7    8    9   10
             ---- ---- ---- ---- ---- ---- ---- ---- ---- ----
  1   1      2.46 0.26 0.39 0.34 1.08 0.56 1.27 1.34 0.29 0.69
  2   2      0.26 4.59 0.39 0.45 0.39 0.75 1.04 0.60 0.78 0.52
  3   3      0.39 0.39 2.91 1.57 0.49 0.45 0.20 1.45 1.17 0.98
  4   4      0.56 0.89 1.68 0.60 0.56 1.66 1.01 1.28 1.34 0.45
  5   5      1.08 0.39 0.49 0.56 2.01 1.12 0.69 1.79 0.39 1.27
  6   6      0.56 0.75 0.45 1.53 1.12 3.58 1.34 0.64 1.34 0.34
  7   7      1.27 1.04 0.20 0.89 0.69 1.34 3.80 1.79 1.37 0.69
  8   8      1.34 0.60 1.45 1.02 1.79 0.64 1.79 2.09 0.78 0.56
  9   9      0.20 0.65 1.17 1.23 0.39 1.23 1.76 0.67 2.01 1.27
 10  10      0.59 0.26 0.88 0.56 1.27 0.34 0.20 0.45 0.88 1.12
```

This part of the output doesn't require much explanation. The expected counts under the null hypothesis of independence are displayed in the top panel and the ratio of observed to expected counts in the second panel. These are typical components of a chi-square based measure of association.

Figure 5.5.3. UCINET Output for Testing Acquaintanceship Differences among Work-Groups (Part 3)



```
Average permutation frequency table

             1     2     3     4     5     6     7     8     9    10
         ----- ----- ----- ----- ----- ----- ----- ----- ----- -----
    1     9.00  7.68 10.23  8.92 10.21  8.99 10.24  8.98 10.23 10.23
    2     7.68  4.80  7.66  6.70  7.58  6.71  7.66  6.75  7.64  7.66
    3    10.21  7.65  8.88  8.97 10.18  8.96 10.21  8.91 10.24 10.22
    4     8.93  6.68  8.97  6.69  8.92  7.91  8.89  7.79  8.95  8.98
    5    10.21  7.57 10.19  8.92  8.86  8.93 10.22  8.96 10.25 10.15
    6     8.98  6.70  8.96  7.89  8.93  6.74  8.92  7.80  8.96  8.96
    7    10.25  7.64 10.19  8.88 10.20  8.94  8.93  8.97 10.22 10.26
    8     8.98  6.74  8.92  7.79  8.95  7.82  8.98  6.67  8.92  8.95
    9    10.23  7.61 10.23  8.94 10.24  8.95 10.22  8.92  8.82 10.23
   10    10.23  7.64 10.25  8.97 10.17  8.96 10.26  8.95 10.23  8.93

Observed chisquare value = 443.100
Significance = 0.000100
Number of iterations = 10000
```

The penultimate bit of the output shows the expected counts under independence – that is, the mean counts observed across 10,000 runs with random assignment of students to workgroups. The differences between the observed and expected counts form a chi-square statistic (443.1). Using the sampling distribution from the permutation experiments, chi-squares this large are observed about one time in 10,000 for random trials.

So, we have strong evidence that there is an association between which work group a student is in (the nodal variable), and the probability that they are acquainted with a student sharing the same attribute (the dyadic variable). This result might be quite sufficient for many questions. But, we might want to go a bit further and explore the actual form of the association. ANOVA density models or "block" models are useful tools for this task.

5.2.3 Multiple Groups: ANOVA Density or "Block" Models

Under the homophily hypothesis, two students who have the same score on an attribute (in this case, being affiliated with a particular work group) may be more likely to know each other (have a dyadic relationship) than two students who do not have the same attribute. If this is the case, then the density of ties among the students in the same work group (regardless of which group it is) ought to be higher than the density of ties of students between students who are not in the same work group.

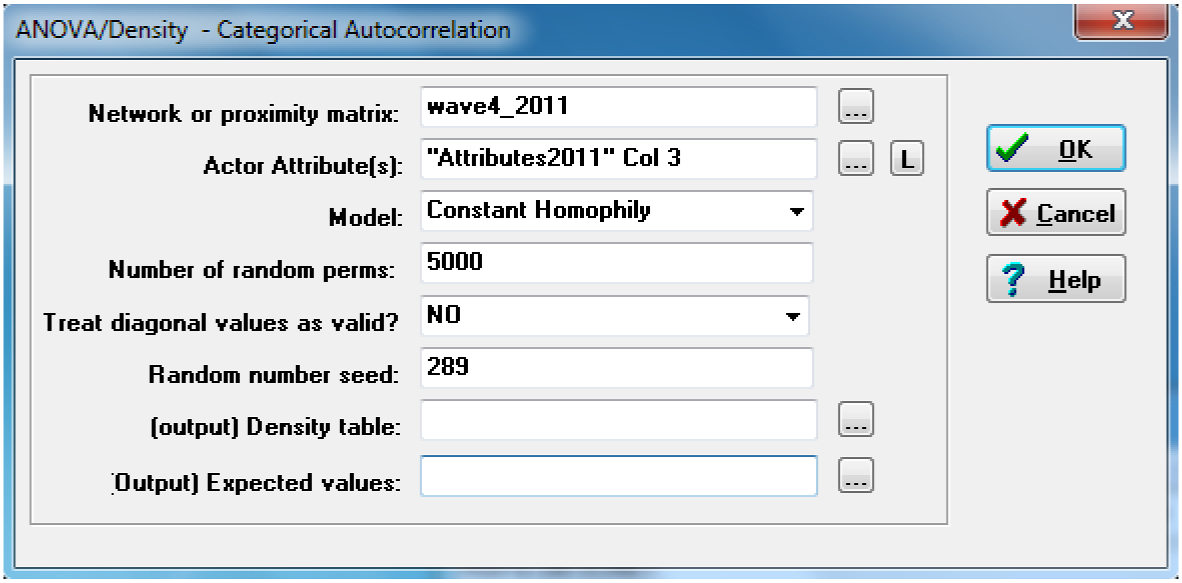
Imagine that we take our original student-by-student acquaintanceship matrix and re-arrange (permute) the rows and columns so that all the students in workgroup 1 are together, followed by workgroup 2, etc. We've "blocked" the matrix according to the "partition" of work-group. The blocked matrix is 10 by 10 groups.

Next, let's count up all of the ties that exist in the blocks that fall along the main diagonal (that is, the block of group 1 with group 1, group 2 with group 2, etc.), and divide them by the number of possible ties. This is the mean, or probability that two actors who are in the same workgroup are acquainted.

Then, let's count up all the ties between pairs of students who are not in the same work groups, and express this as proportion or mean of all possible ties that could have existed among these students.

We now have two groups (dyads in the same work group; dyads not in the same work group), and we can perform a test of differences of means, or one-way ANOVA. Figure 5.6 shows how to perform this particular two-group, one-way, ANOVA with UCINET's *Tools>Testing Hypotheses>Mixed Dyadic/Nodal>Categorical Attribute>ANOVA Density.*

Figure 5.6. UCINET Dialog for Testing Acquaintanceship by Workgroup using a Constant Homophily Block Model


ANOVA/Density - Categorical Autocorrelation

Network or proximity matrix: wave4_2011

Actor Attribute(s): "Attributes2011" Col 3

Model: Constant Homophily

Number of random perms: 5000

Treat diagonal values as valid? NO

Random number seed: 289

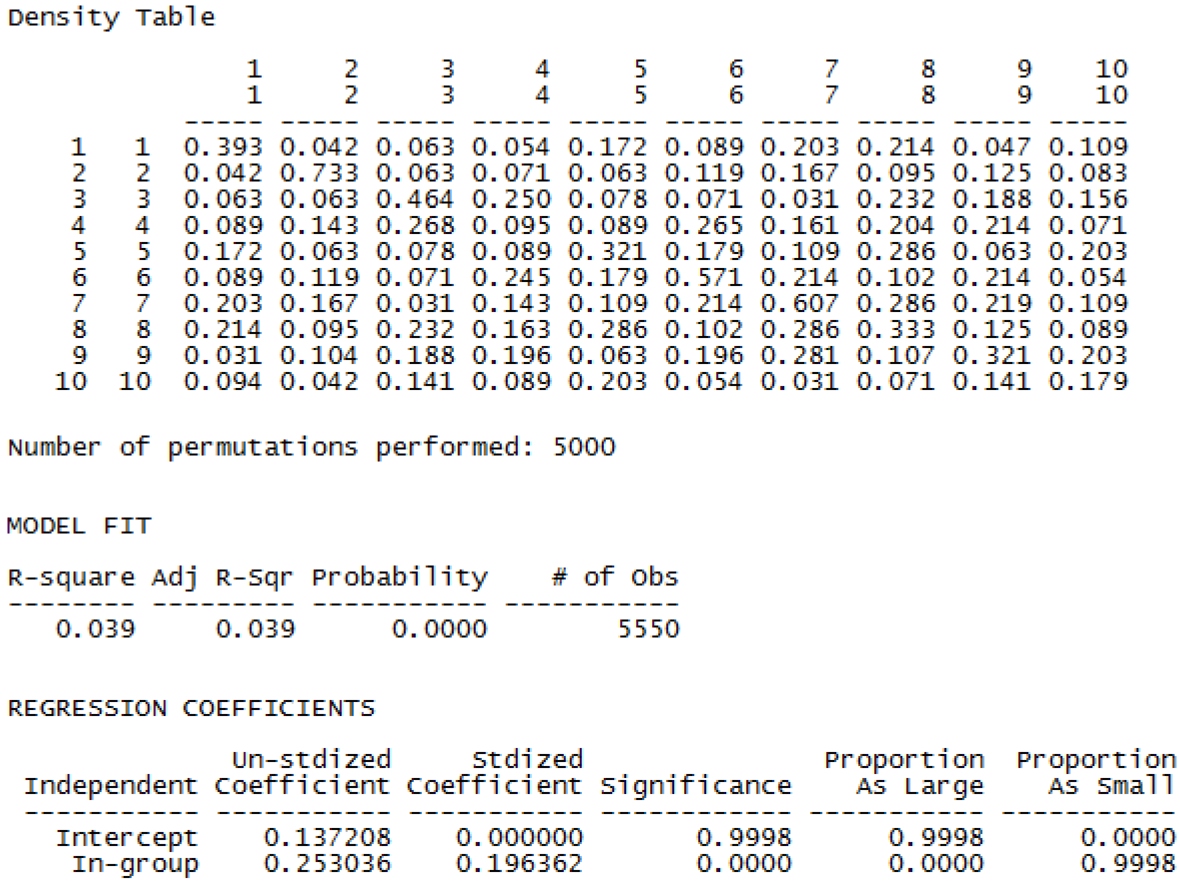
(output) Density table:

(Output) Expected values:

As always, there are controls for the permutation tests, and for saving parts of the output as UCINET files for further processing. We select our acquaintanceship dyadic data (it's being treated here as the dependent variable), and the vector of work group affiliation from the attributes file (it happens to be the third column in that file).

The important part is the "Model," where we've chosen "Constant Homophily." The hypothesis that we have been discussing suggests that there is a difference in the likelihood of a tie between two actors in the same workgroup from the likelihood of a tie between two actors who are not in the same workgroup. Returning to our acquaintanceship matrix that has been blocked by workgroup, we are saying that the mean densities of all the blocks on the main diagonal (i.e. densities of ties to others in the same workgroup) are the same, and that the densities of all of the blocks not on the main diagonal (i.e. densities of ties to others outside of one's group) are the same – and that these two densities differ. That is, there is a tendency toward homophily (the diagonal blocks differ from the non-diagonal blocks), and this tendency is constant across work groups (the density of ties within workgroup 1 is the same as the density of ties within workgroup 2, etc.). If we expected the tendency toward homophily to vary by work group, we could choose "Variable Homophily" instead. If we were interested in tendencies for out-group tie formation (vs. the in-group tendencies assumed in homophily assumptions), we would choose "Structural Blockmodel." Figure 5.7 shows the output for our constant homophily model.

Figure 5.7. UCINET Output for Testing Acquaintanceship by Workgroup using a Constant Homophily Block Model



```
Density Table

                  1     2     3     4     5     6     7     8     9    10
                  1     2     3     4     5     6     7     8     9    10
              ----- ----- ----- ----- ----- ----- ----- ----- ----- -----
    1    1    0.393 0.042 0.063 0.054 0.172 0.089 0.203 0.214 0.047 0.109
    2    2    0.042 0.733 0.063 0.071 0.063 0.119 0.167 0.095 0.125 0.083
    3    3    0.063 0.063 0.464 0.250 0.078 0.071 0.031 0.232 0.188 0.156
    4    4    0.089 0.143 0.268 0.095 0.089 0.265 0.161 0.204 0.214 0.071
    5    5    0.172 0.063 0.078 0.089 0.321 0.179 0.109 0.286 0.063 0.203
    6    6    0.089 0.119 0.071 0.245 0.179 0.571 0.214 0.102 0.214 0.054
    7    7    0.203 0.167 0.031 0.143 0.109 0.214 0.607 0.286 0.219 0.109
    8    8    0.214 0.095 0.232 0.163 0.286 0.102 0.286 0.333 0.125 0.089
    9    9    0.031 0.104 0.188 0.196 0.063 0.196 0.281 0.107 0.321 0.203
   10   10    0.094 0.042 0.141 0.089 0.203 0.054 0.031 0.071 0.141 0.179

Number of permutations performed: 5000

MODEL FIT

R-square Adj R-Sqr Probability    # of Obs
-------- --------- ----------- -----------
   0.039     0.039      0.0000        5550

REGRESSION COEFFICIENTS

               Un-stdized      Stdized                  Proportion  Proportion
 Independent  Coefficient  Coefficient  Significance    As Large    As Small
 -----------  -----------  -----------  ------------  ----------- -----------
   Intercept     0.137208     0.000000        0.9998       0.9998      0.0000
    In-group     0.253036     0.196362        0.0000       0.0000      0.9998
```

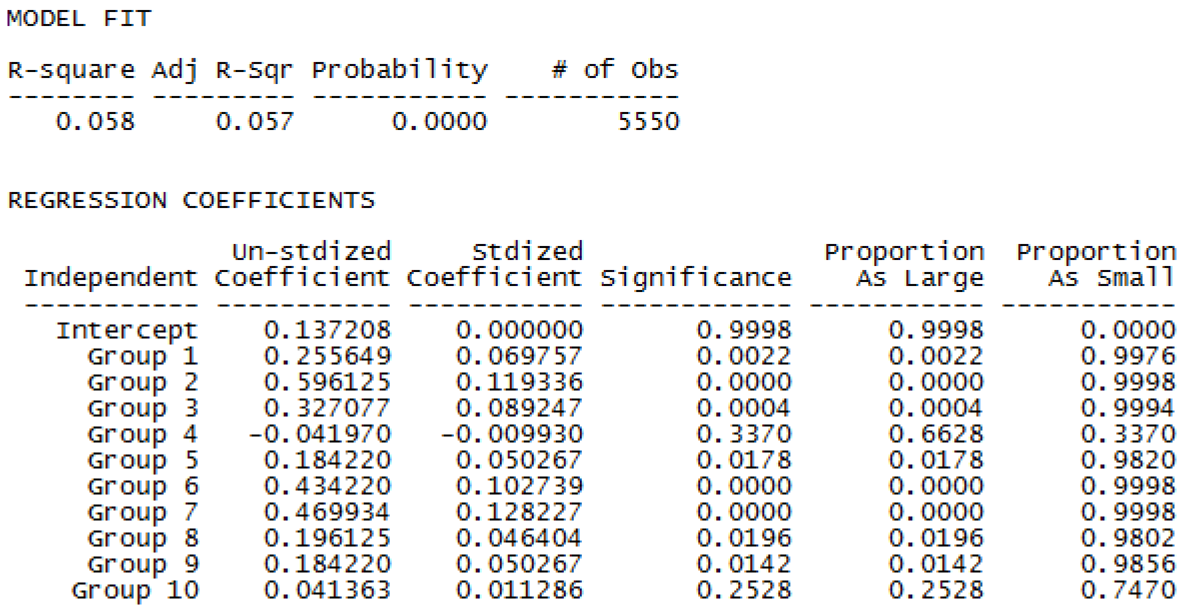
The first portion of the output shows the block densities actually present in our data. That is, it shows the mean density (or with a binary variable, the probability) of a tie between any two actors within a given block. For example, two students in work-group 1 have a 0.393 chance of having a tie, while the probability of a tie from a student in group 1 to group 3 is only 0.063. Just looking at the data, we see that the densities on the main diagonal (homophilic ties) are generally higher than those off of the main diagonal. We also should note that the densities of the blocks on the main diagonal are not very “constant” or similar to one another. Different work-groups have rather different degrees of homophily.

The next panels of the output describe how well the constant homophily block model does in explaining or predicting ties between pairs of actors. The answer is: not very well. Classifying dyads as being either within the same work group or not in the same workgroup explains about 4% of the variance in the probability that there is a dyadic tie. Note, however, that the permutation trials test very strongly suggests that workgroup homophily does have an effect ($p < 0.0001$).

The last part of the output shows, and tests, specific mean differences as regression coefficients (or effects in ANOVA language). We see that the mean probability of a tie between two students who are not in the same workgroup (the reference category or intercept) is 0.137. The probability of a tie between two students who are in the same work group is 0.253 higher, or 0.39. We see that this is a significant difference. Even though the variance explained is not high, being in the same work group more than doubles the probability of a dyadic tie between two students.

The constant homophily block model expresses a pretty strong hypothesis that suggests that there are no meaningful differences across workgroups in their tendency toward in-group tie formation. It also suggests that the bias against out-group acquaintances are the same, regardless of the out-group. An alternative model might propose that the bias against ties outside ones group are the same for all outsiders, but that groups differ in their tendency toward in-group ties. This hypothesis is selected by choosing the “variable homophily” block model in the dialog (figure 5.6). The output for this model is shown in figure 5.8.

Figure 5.8. UCINET Output for Testing Acquaintanceship by Workgroup using a Variable Homophily Block Model



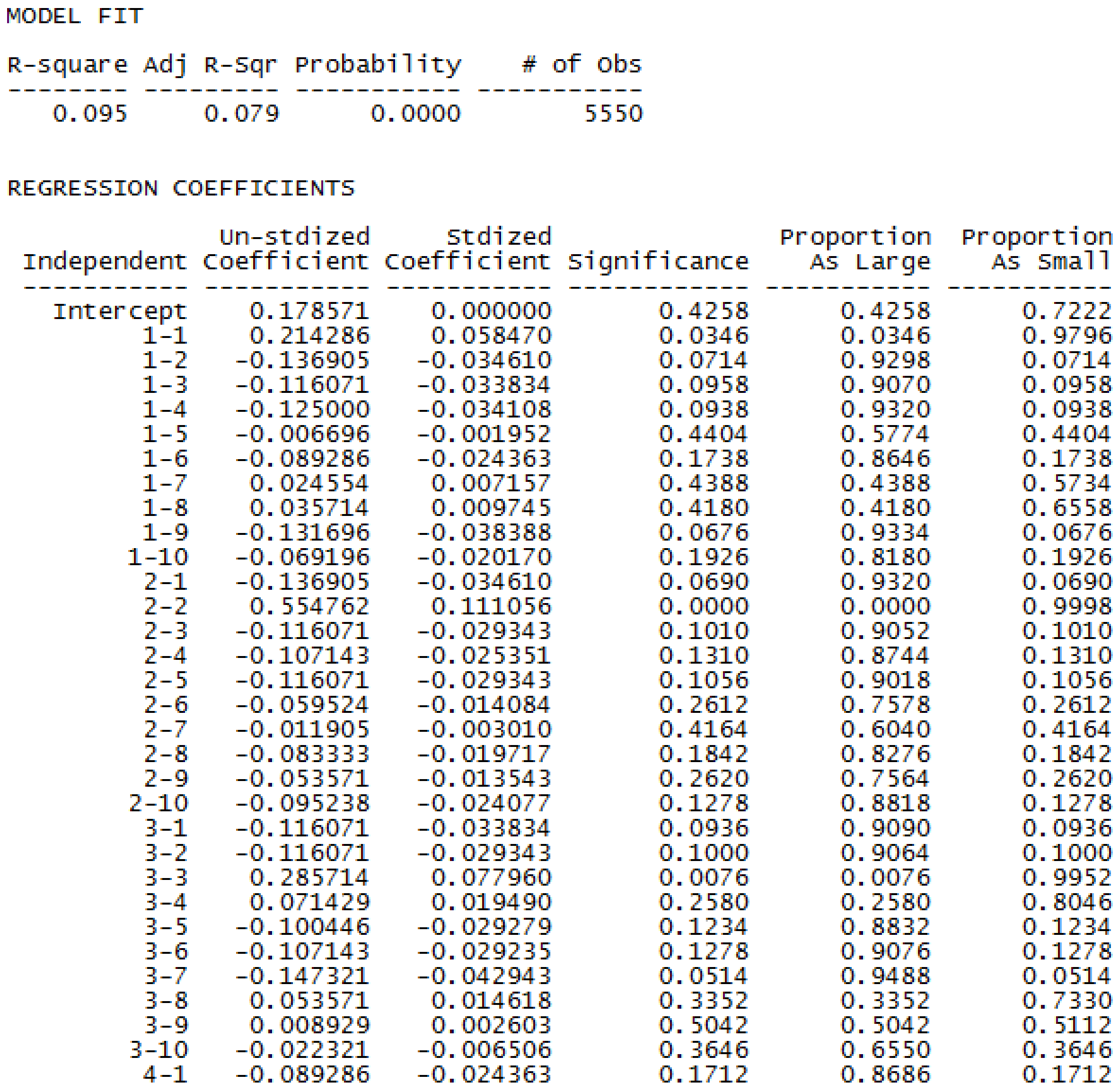
```
MODEL FIT

R-square Adj R-Sqr Probability    # of Obs
-------- --------- ----------- -----------
   0.058     0.057      0.0000        5550

REGRESSION COEFFICIENTS

              Un-stdized      Stdized                   Proportion  Proportion
 Independent Coefficient Coefficient Significance     As Large    As Small
 ----------- ----------- ----------- ------------ ----------- -----------
   Intercept    0.137208    0.000000       0.9998      0.9998      0.0000
     Group 1    0.255649    0.069757       0.0022      0.0022      0.9976
     Group 2    0.596125    0.119336       0.0000      0.0000      0.9998
     Group 3    0.327077    0.089247       0.0004      0.0004      0.9994
     Group 4   -0.041970   -0.009930       0.3370      0.6628      0.3370
     Group 5    0.184220    0.050267       0.0178      0.0178      0.9820
     Group 6    0.434220    0.102739       0.0000      0.0000      0.9998
     Group 7    0.469934    0.128227       0.0000      0.0000      0.9998
     Group 8    0.196125    0.046404       0.0196      0.0196      0.9802
     Group 9    0.184220    0.050267       0.0142      0.0142      0.9856
    Group 10    0.041363    0.011286       0.2528      0.2528      0.7470
```

We see that the variable homophily model provides an improvement in fit (R-squared is 0.058, up from 0.039). The result is very unlikely to arise from random processes, but the model does not account for very much of the variation in who is acquainted with whom.

This improvement in fit has been bought at the expense of 9 additional parameters. Now, the tendency toward homophily in each of the 10 workgroups is allowed to differ from the intercept (which is the probability of a tie with an out-group student). The groups vary considerably in their tendency toward internal ties, as we see from the slopes describing how the density within each significantly differs from the intercept (group 4 and group 10 being exceptions).

The variable homophily block model allows for differences among groups in their preference for in-group ties. It treats all ties outside one's own group as homogeneous. We can take the final step of relaxing this assumption with the "structural block model" option in the ANOVA density models dialog. Only a portion of the lengthy output is shown in figures 5.9.

Figure 5.9. UCINET Output for Testing Acquaintanceship by Workgroup using a Structural Block Model



```
MODEL FIT

R-square Adj R-Sqr Probability    # of Obs
-------- --------- ----------- -----------
   0.095     0.079      0.0000        5550


REGRESSION COEFFICIENTS

              Un-stdized      Stdized              Proportion  Proportion
 Independent Coefficient Coefficient Significance    As Large    As Small
 ----------- ----------- ----------- ------------ ----------- -----------
   Intercept    0.178571    0.000000       0.4258      0.4258      0.7222
         1-1    0.214286    0.058470       0.0346      0.0346      0.9796
         1-2   -0.136905   -0.034610       0.0714      0.9298      0.0714
         1-3   -0.116071   -0.033834       0.0958      0.9070      0.0958
         1-4   -0.125000   -0.034108       0.0938      0.9320      0.0938
         1-5   -0.006696   -0.001952       0.4404      0.5774      0.4404
         1-6   -0.089286   -0.024363       0.1738      0.8646      0.1738
         1-7    0.024554    0.007157       0.4388      0.4388      0.5734
         1-8    0.035714    0.009745       0.4180      0.4180      0.6558
         1-9   -0.131696   -0.038388       0.0676      0.9334      0.0676
        1-10   -0.069196   -0.020170       0.1926      0.8180      0.1926
         2-1   -0.136905   -0.034610       0.0690      0.9320      0.0690
         2-2    0.554762    0.111056       0.0000      0.0000      0.9998
         2-3   -0.116071   -0.029343       0.1010      0.9052      0.1010
         2-4   -0.107143   -0.025351       0.1310      0.8744      0.1310
         2-5   -0.116071   -0.029343       0.1056      0.9018      0.1056
         2-6   -0.059524   -0.014084       0.2612      0.7578      0.2612
         2-7   -0.011905   -0.003010       0.4164      0.6040      0.4164
         2-8   -0.083333   -0.019717       0.1842      0.8276      0.1842
         2-9   -0.053571   -0.013543       0.2620      0.7564      0.2620
        2-10   -0.095238   -0.024077       0.1278      0.8818      0.1278
         3-1   -0.116071   -0.033834       0.0936      0.9090      0.0936
         3-2   -0.116071   -0.029343       0.1000      0.9064      0.1000
         3-3    0.285714    0.077960       0.0076      0.0076      0.9952
         3-4    0.071429    0.019490       0.2580      0.2580      0.8046
         3-5   -0.100446   -0.029279       0.1234      0.8832      0.1234
         3-6   -0.107143   -0.029235       0.1278      0.9076      0.1278
         3-7   -0.147321   -0.042943       0.0514      0.9488      0.0514
         3-8    0.053571    0.014618       0.3352      0.3352      0.7330
         3-9    0.008929    0.002603       0.5042      0.5042      0.5112
        3-10   -0.022321   -0.006506       0.3646      0.6550      0.3646
         4-1   -0.089286   -0.024363       0.1712      0.8686      0.1712
```

The structural block model allows each of the 10 by 10 blocks to vary in density. The variance explained by allowing variation in the external tie densities is relatively substantial (up to 9.5% from 5.8%). But, the gain is bought at the expense of 99 degrees of freedom. Every one of the 10x10 blocks is deviated from the last block (group 10 ties with group 10).

The ANOVA density, or “block” model allows fitting a variety of hypotheses about how a categorical attribute partitions (or blocks, or relates to) a relational (dyadic) variable. The relational variable can be either binary, as in our example, or continuous. For example, the relational variable might express the strength of ties in dyads, or it might express the network distance between two actors. So, block modeling can be used with categorical attributes and either categorical or continuous relational variables.

5.3 Continuous Attributes

In looking at the relationship between an attribute and a network, when the attribute of interest is continuous, spatial autocorrelation methods can be used. These methods are borrowed from geographical analysis, and are adapted to use "social distance" between actors in a network, rather than spatial distances.

The type of question that we are addressing here is whether actors who are closer to one another in a network are likely to have the same score on an attribute. For example, we might ask whether students who know one another have similar exam scores. Questions like this one lie at the core of social influence theory, where it is hypothesized that the stronger the ties are between two actors, the more likely they are to converge in attitudes and behaviors as they model on and influence one another.

The core idea of network autocorrelation is that the correlation between the attributes of two members of a dyad covaries with the distance or strength of the tie between them. The basic method for assessing this is, logically, pretty simple. For each dyad, the individual nodal scores of both actors on some attribute are recorded (for example, the final exam scores of AD and AJ). Some measure of the strength of the tie or of the network distance between the two actors is generated (for example, say that the geodesic distance from AD to AJ is two steps of acquaintanceship, details below). The scores on the attribute of the two members of the dyad are then correlated across dyads, but weighted according to closeness or strength of the tie between them. The result is a distance-weighted correlation that tells us whether two actors who are close together are likely to have similar scores on the attribute (positive autocorrelation, most common) or are likely to have dissimilar scores on the attribute (negative autocorrelation), or if distance is irrelevant to the correlation between the scores of the members of dyads.

To examine autocorrelation, we need the attribute of interest for each actor. This is simply the attribute vector (e.g. test scores). We also need a measure of the distance between the two actors, so that we can "weight" the data. These weights are a dyadic variable describing the distance or strength of the tie between the members of the dyad.

In spatial autocorrelation applications (from which SNA appropriated these methods), the distance between two nodes (say locations on a map) is a fairly straightforward idea (actually, geo-spatial approaches to distance can be quite subtle, interesting, and complex). But, how do we measure "social distance?" There is no single best answer, but consider some possibilities.

One possible measure of dyadic distance for autocorrelation weighting is social similarity (such as distance in "Blau-space"; McPherson & Ranger-Moore, 1991). A dyadic variable could be created that measures similarity across a variety of attributes, or similarity in patterns of affiliation, for each pair of actors. Scaling, clustering, and index construction methods might then be used to calculate a "similarity" matrix for dyads.

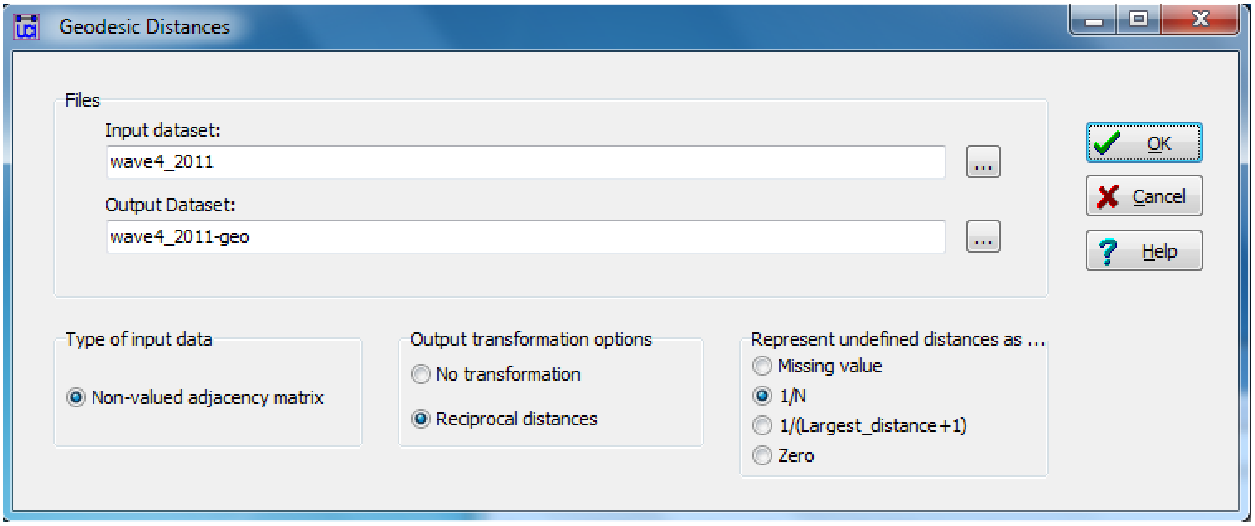
Another, more concrete approach to creating distance weights is to see how far apart two actors are in a social network, directly. If AB names AJ as an acquaintance, the distance between them is 1 step. If AB names AJ and AJ names BL (and AB does not name BL), then the distance between AB and BL is 2 steps. Geodesic distance (the length of the shortest path from one actor to another) is a common approach and can be directly generated by UCINET (*Network>Cohesion>Geodesic Distances*).

Of course, the adjacency matrix can also be used directly. In this case, the distance between two actors in a network is either zero or one. Or, if the ties were measured on a continuous scale of strength, the matrix of tie strengths can be used directly.

For our example, we'll define the distance between two students as the length of the geodesic path between them. But, since we would like a positive autocorrelation to mean that similarity increases with nearness (not distance), we will need to reverse the direction of the weights. We could simply subtract each distance from the maximum distance to measure closeness rather than distance. But, we are going to take the reciprocal of the geodesic distance instead. The reciprocal of the geodesic distance not only indexes "nearness," but it also scales the data so that influence declines at an accelerated rate as distance increases (adjacent actors get a weight of 1/1 or 1.0; actors at a distance of two get a weight of 1/2 or 0.5, etc.).

Figure 5.10 shows the dialog for generating nearness (inverse distance) weights for our acquaintanceship data at the time of the final exam in UCINET by asking for geodesic distances (*UCINET>Network>Cohesion>Geodesic Distances*), rescaled by taking the reciprocal.

Figure 5.10. UCINET Dialog Used to Generate Inverse Geodesic Distances for Wave 4 Directed Acquaintanceship



Geodesic Distances

Files

Input dataset: wave4_2011

Output Dataset: wave4_2011-geo

Type of input data

● Non-valued adjacency matrix

Output transformation options

○ No transformation

● Reciprocal distances

Represent undefined distances as ...

○ Missing value

● 1/N

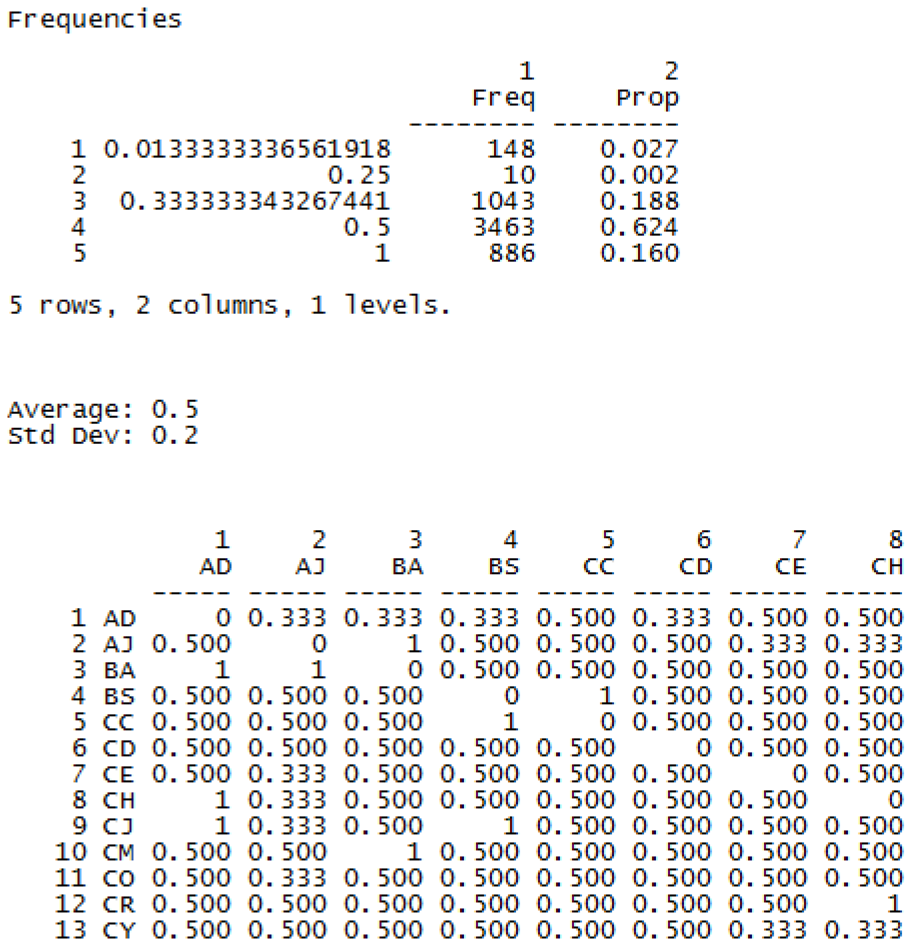
○ 1/(Largest_distance+1)

○ Zero

We can save the dyadic nearness variable that we're creating as a new dataset and take the default name. Notice that we've marked the circle "1/N" in order to produce reciprocal distance values.

A portion of the nearness weights matrix is shown in Figure 5.11. We can also see from the figure that the most frequently occurring nearness weight is 0.5 which corresponds to a degree of 2.

Figure 5.11. UCINET Output for the Inverse Geodesic Distances of Wave 4 Directed Acquaintanceship

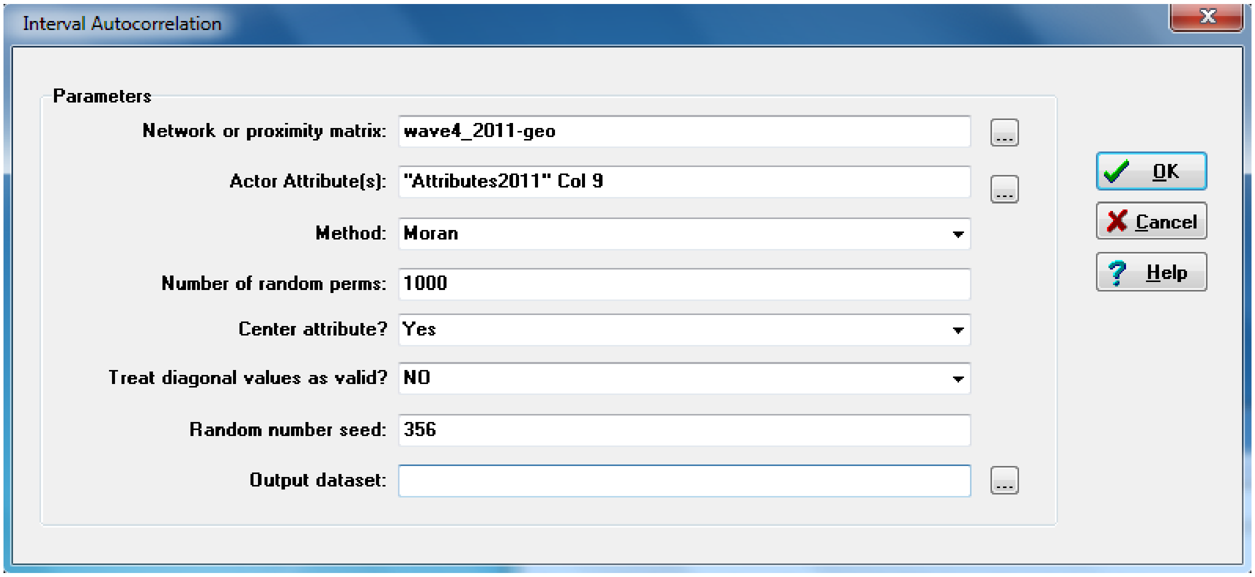


```
Frequencies

                               1        2
                            Freq     Prop
                        -------- --------
   1 0.0133333336561918      148    0.027
   2               0.25       10    0.002
   3  0.333333343267441     1043    0.188
   4                0.5     3463    0.624
   5                  1      886    0.160

5 rows, 2 columns, 1 levels.


Average: 0.5
Std Dev: 0.2


                1     2     3     4     5     6     7     8
               AD    AJ    BA    BS    CC    CD    CE    CH
            ----- ----- ----- ----- ----- ----- ----- -----
    1 AD        0 0.333 0.333 0.333 0.500 0.333 0.500 0.500
    2 AJ    0.500     0     1 0.500 0.500 0.500 0.333 0.333
    3 BA        1     1     0 0.500 0.500 0.500 0.500 0.500
    4 BS    0.500 0.500 0.500     0     1 0.500 0.500 0.500
    5 CC    0.500 0.500 0.500     1     0 0.500 0.500 0.500
    6 CD    0.500 0.500 0.500 0.500 0.500     0 0.500 0.500
    7 CE    0.500 0.333 0.500 0.500 0.500 0.500     0 0.500
    8 CH        1 0.333 0.500 0.500 0.500 0.500 0.500     0
    9 CJ        1 0.333 0.500     1 0.500 0.500 0.500 0.500
   10 CM    0.500 0.500     1 0.500 0.500 0.500 0.500 0.500
   11 CO    0.500 0.333 0.500 0.500 0.500 0.500 0.500 0.500
   12 CR    0.500 0.500 0.500 0.500 0.500 0.500 0.500     1
   13 CY    0.500 0.500 0.500 0.500 0.500 0.500 0.333 0.333
```

For network autocorrelation methods, a distance weighting matrix is needed. It is a dyadic variable that measures how similar, strong, or close the tie is between each pair of actors. As you can see, measuring the social distance between the two members of a dyad might be done in a wide variety of ways. And, as you can readily imagine, how one defines distance and creates a nearness weighting matrix can dramatically affect the results of an analysis. Our example is only one possible approach.

Once we have decided what attribute we want to test for network autocorrelation, and once we have created a distance or nearness weighting matrix, it is quite simple to generate some standard measures of network autocorrelation in UCINET (*UCINET>Tools>Testing Hypotheses>Mixed Dyadic/Nodal>Continuous Attributes>Moran/Geary statistics*). Figure 5.12 shows the dialog.

Figure 5.12. UCINET Dialog for Network Autocorrelation



Interval Autocorrelation

Parameters

Network or proximity matrix: wave4_2011-geo

Actor Attribute(s): "Attributes2011" Col 9

Method: Moran

Number of random perms: 1000

Center attribute? Yes

Treat diagonal values as valid? NO

Random number seed: 356

Output dataset:

In the first box of the dialog, the nearness weights dyadic variable is identified. Note that any dyadic variable could be used – measures of social similarity, joint affiliation, as well as geographic or network distance. In the second box of the dialog the attribute that we want to examine for autocorrelation is identified. Here we identify column nine of the attribute dataset: student's grades on the final exam (an interval/ratio variable). Distance (or nearness) weighted autocorrelations can, of course, be calculated for binary or ordinal-scale attributes, but the correlation measures were intended for continuous attributes.

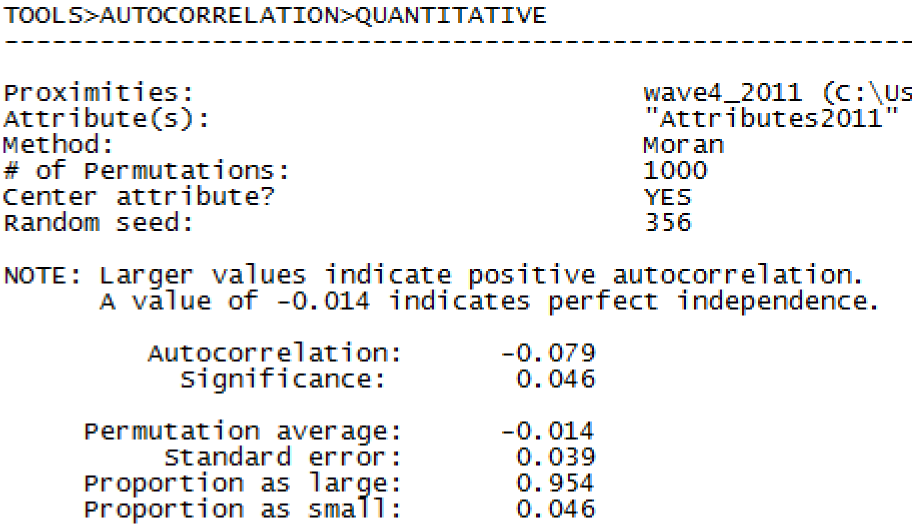
5.3.1 Global Network Autocorrelation

There are quite a number of measures of spatial autocorrelation that have been developed, primarily by geo-scientists (Getis and Ord, 1992; Griffith, 1987). UCINET provides the two most commonly used: Moran and Geary indexes. The Moran index (and some others, as we will see below) summarizes the global pattern of autocorrelation over the entire network.

That is, it looks across the whole network, seeking broad patterns. In our example using final exam grades, the Moran index is “looking” for big regions of the network that are occupied by mostly high scoring students and other big regions, or network neighborhoods, or communities that are composed mostly of low scoring students. The Moran index is somewhat less sensitive to an individual’s immediate neighborhoods.

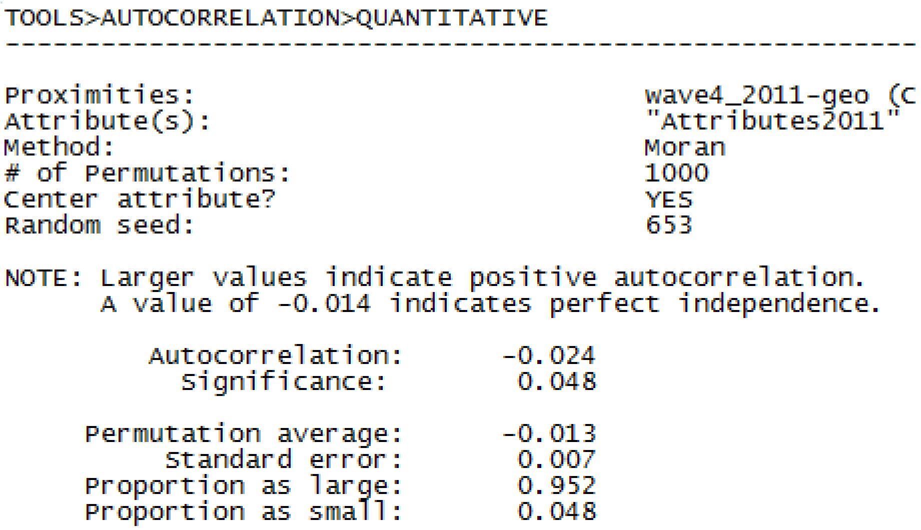
In figures 5.13 and 5.14 the output of *Tools>Testing Hypotheses>Mixed Dyadic/Nodal>Continuous Attributes>Moran Geary Statistics* is shown (selecting the Moran statistic) using two different measures of nearness. In the first run (shown in figure 5.13), we use simple adjacency (A is acquainted with B, or not) as our measure of the network closeness or nearness weight. The “network or proximity matrix” used to generate the output in figure 5.13 is our wave 4 acquaintance data (“wave4_2011”). In the second run (figure 5.14), we use the reciprocal of the geodesic distance as our nearness weights (“wave4_2011-geo”). The dialog box shown in figure 5.12 generates the output shown in figure 5.14.

Figure 5.13. UCINET Output for Network Adjacency Autocorrelation of Final Exam Grades (Moran Index)



```
TOOLS>AUTOCORRELATION>QUANTITATIVE
--------------------------------------------------------

Proximities:                           wave4_2011 (C:\Us
Attribute(s):                          "Attributes2011"
Method:                                Moran
# of Permutations:                     1000
Center attribute?                      YES
Random seed:                           356

NOTE: Larger values indicate positive autocorrelation.
      A value of -0.014 indicates perfect independence.

         Autocorrelation:      -0.079
           Significance:        0.046

     Permutation average:      -0.014
          Standard error:       0.039
     Proportion as large:       0.954
     Proportion as small:       0.046
```

Figure 5.14. UCINET Output for Network Inverse Geodesic Distance Autocorrelation of Final Exam Grades (Moran Index)



```
TOOLS>AUTOCORRELATION>QUANTITATIVE
--------------------------------------------------------

Proximities:                              wave4_2011-geo (C
Attribute(s):                             "Attributes2011"
Method:                                   Moran
# of Permutations:                        1000
Center attribute?                         YES
Random seed:                              653

NOTE: Larger values indicate positive autocorrelation.
      A value of -0.014 indicates perfect independence.

         Autocorrelation:      -0.024
           Significance:        0.048

     Permutation average:      -0.013
          Standard error:       0.007
     Proportion as large:       0.952
     Proportion as small:       0.048
```

Interpreting the Moran statistic is a bit tricky in our case. Larger negative values of the Moran statistic indicate positive autocorrelation (here, a tendency for students with similar final exam scores to be adjacent or at shorter distances in the network) because we are using nearness weights rather than the typical use of distance. Significance tests are performed, as usual, using permutation (randomly shuffling the attribute scores).

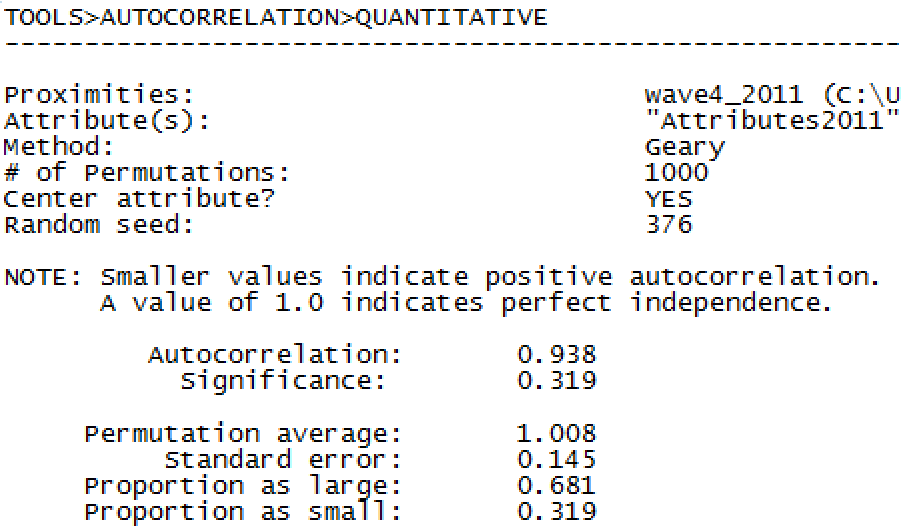
When nearness is defined strictly as adjacency (Fig. 5.13), we see a modest positive autocorrelation (Moran statistic = -0.079) that is significant at the p = 0.05 level. When we apply inverse distance weights (Fig. 5.14), including indirect alters, the positive autocorrelation is weaker (-0.024).

In either case, we see a weak, but probably non-random, tendency for the network as a whole to display communities, regions, or neighborhoods occupied by students with similar grades on the final exam. There is some evidence that knowing one another and having similar grades are weakly associated.

5.3.2 Local Network Autocorrelation

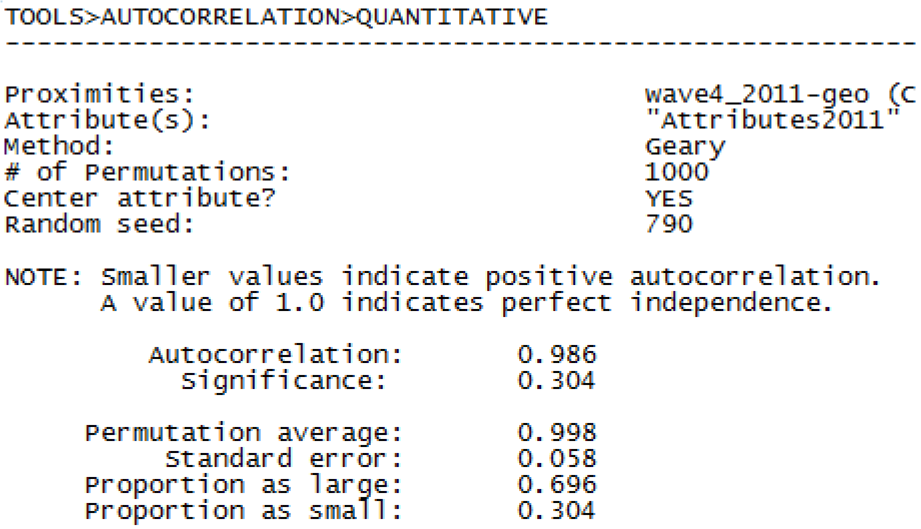
The *Geary* statistic is constructed somewhat differently from the Moran, and places a greater emphasis on local patterns in the network. Rather than focusing on the network as a whole and looking for large regions of similar actors, it focuses on each actor's local ties and then aggregates local autocorrelations into a global average. Figures 5.15 and 5.16 show the Geary autocorrelations of student's grades where nearness is defined as adjacency (figure 5.15) or the reciprocal of geodesic distance (figure 5.16).

5.15. UCINET Output for Network Adjacency Autocorrelation of Final Exam Grades (Geary Index)



```
TOOLS>AUTOCORRELATION>QUANTITATIVE
--------------------------------------------------------

Proximities:                            wave4_2011 (C:\U
Attribute(s):                           "Attributes2011"
Method:                                 Geary
# of Permutations:                      1000
Center attribute?                       YES
Random seed:                            376

NOTE: Smaller values indicate positive autocorrelation.
      A value of 1.0 indicates perfect independence.

         Autocorrelation:       0.938
           Significance:        0.319

     Permutation average:       1.008
          Standard error:       0.145
     Proportion as large:       0.681
     Proportion as small:       0.319
```

Figure 5.16. UCINET Output for Network Inverse Geodesic Distance Autocorrelation of Final Exam Grades (Geary Index)



```
TOOLS>AUTOCORRELATION>QUANTITATIVE
---------------------------------------------------------

Proximities:                            wave4_2011-geo (C
Attribute(s):                           "Attributes2011"
Method:                                 Geary
# of Permutations:                      1000
Center attribute?                       YES
Random seed:                            790

NOTE: Smaller values indicate positive autocorrelation.
      A value of 1.0 indicates perfect independence.

         Autocorrelation:       0.986
           Significance:        0.304

     Permutation average:       0.998
          Standard error:       0.058
     Proportion as large:       0.696
     Proportion as small:       0.304
```

The Geary statistic is a somewhat odd metric. With Geary, more positive values indicate larger positive autocorrelations because we are using nearness matrices rather than the typical distance matrices used in geospatial analyses. However, the statistic ranges from zero, indicating perfect negative autocorrelation, to +2.0, indicating perfect positive autocorrelation. A value of +1.0 indicates no autocorrelation. Our values of 0.938 and 0.986 reported in our results, indicate a very slight negative autocorrelation, on the average, among the neighborhoods of individual students, but the effect is not significant in either case.

The values for the Geary statistics for both models of nearness are in the opposite direction from the Moran values. But, the Geary statistic does not reach statistical significance (at the p = 0.05 level) in either model. As with the Moran statistic, the magnitude of autocorrelation is stronger using simple adjacency, rather than inverse geodesic distances. This may not be surprising for many social influence processes (e.g. what our friend thinks of us is important, but what a friend of a friend thinks of us may not be relevant).

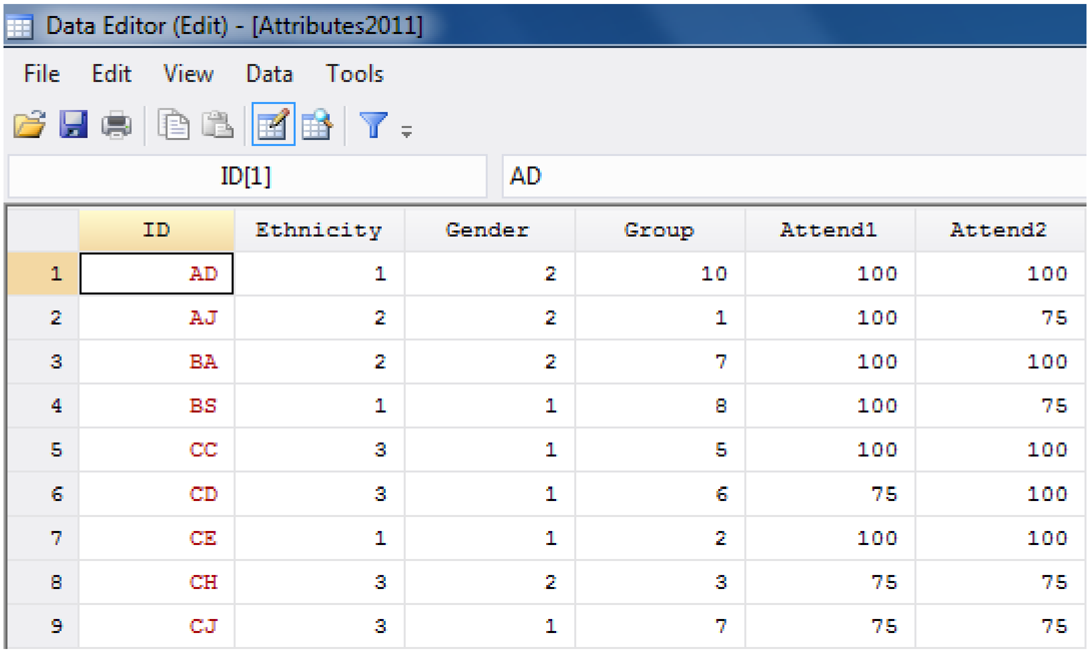
5.3.3 Network Autocorrelation with Stata

UCINET provides quick and easy calculation of the two most commonly used spatial autocorrelation measures, and good tools for constructing distance-weighting dyadic variables. Other software is available that can calculate additional measures of autocorrelation, and more importantly, use autocorrelation in predictive regression models. Maurizio Pisati (2001, 2012) has built a set of routines in Stata that are helpful for these tasks. Here, we'll briefly illustrate how to use these tools for calculating measures of autocorrelation. In the next chapter, we will apply them to regression models that include network autocorrelation.

First, we need to locate and download the toolkit that Pisati has built by doing a *findit* in Stata. The components that are needed are *spatwmat* (which imports a distance weights matrix and calculates the needed eigenvectors), *spatgsa* (which calculates the Moran measure of global spatial autocorrelation), and *spatlsa* (which calculates local versions of the Moran measure for individual neighborhoods). The library "spatreg" (which performs spatial autoregression models) will be used in the next chapter. Stata help files come with the libraries, and can be viewed with the help command (*help spatwmat* for example) after the libraries are installed.

Second, we need to create Stata data files (i.e. ".dta" files) for our attribute (student grades), and social distance weights (we'll use the reciprocal geodesic distances for our illustration). First, make sure that the rows and columns of the distance weights data are in the same sort order, and that the sort order agrees with that of the attributes dataset. Then, one can cut-and-paste, or use exports and imports to create two data sets in Stata. Figures 5.17 and 5.18 show partial screenshots of our two datasets.

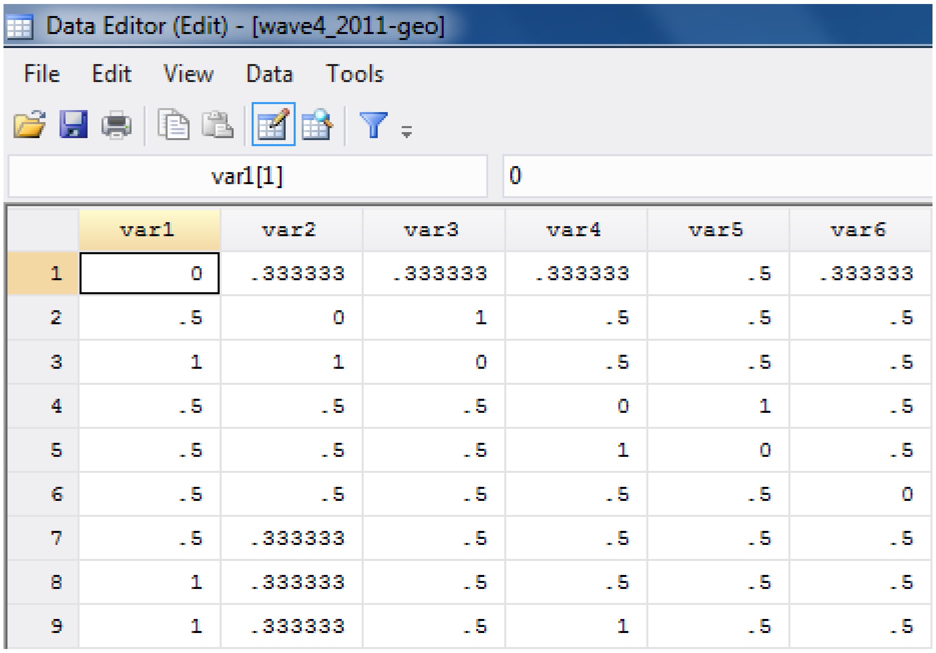
Figure 5.17. Stata Data Editor Displaying Attributes Data Stored as a ".dta" File



ID[1] AD

	ID	Ethnicity	Gender	Group	Attend1	Attend2
1	AD	1	2	10	100	100
2	AJ	2	2	1	100	75
3	BA	2	2	7	100	100
4	BS	1	1	8	100	75
5	CC	3	1	5	100	100
6	CD	3	1	6	75	100
7	CE	1	1	2	100	100
8	CH	3	2	3	75	75
9	CJ	3	1	7	75	75

Figure 5.18. Stata Data Editor Displaying Distance Weights Matrix Stored as a ".dta" File

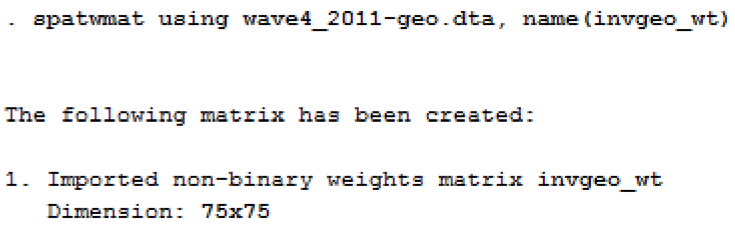


	var1	var2	var3	var4	var5	var6
1	0	.333333	.333333	.333333	.5	.333333
2	.5	0	1	.5	.5	.5
3	1	1	0	.5	.5	.5
4	.5	.5	.5	0	1	.5
5	.5	.5	.5	1	0	.5
6	.5	.5	.5	.5	.5	0
7	.5	.333333	.5	.5	.5	.5
8	1	.333333	.5	.5	.5	.5
9	1	.333333	.5	1	.5	.5

Note that the variable names (names of the nodes) are not included as a variable in the distance file, which has arbitrary variable names.

The third step of the process is to run *spatwmat* to generate the necessary formatted distances and store them in a temporary file in the working directory. Figure 5.19 shows how we did this step.

Figure 5.19. Stata Syntax for Running *spatwmat* Command



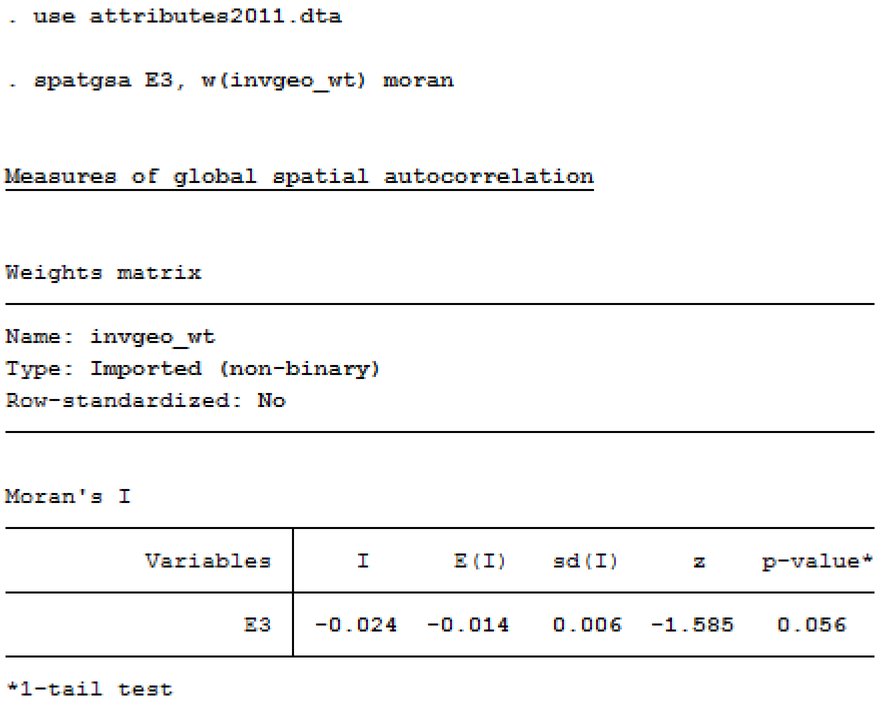
```
. spatwmat using wave4_2011-geo.dta, name(invgeo_wt)


The following matrix has been created:

1. Imported non-binary weights matrix invgeo_wt
   Dimension: 75x75
```

The Stata command tells the processor to run *spatwmat* using the distance weights Stata file we created ("wave4_2011-geo.dta"). After the comma (which delimits options in Stata), is the required option of name(), where you provide a name for the temporary weights file.

We called ours “invgeo_wt.” *Spatwmat* reports that it created a non-binary weights matrix of 75x75.

Finally, we are ready to calculate measures of global autocorrelation for the variable “E3” which is the score on the final examination. Figure 5.20 shows how this is done. Note this must be done in the same working session where we create the temporary data file with the weights!

Figure 5.20. Stata Syntax and Output for Global Autocorrelation of Final Exam Using *spatgsa*

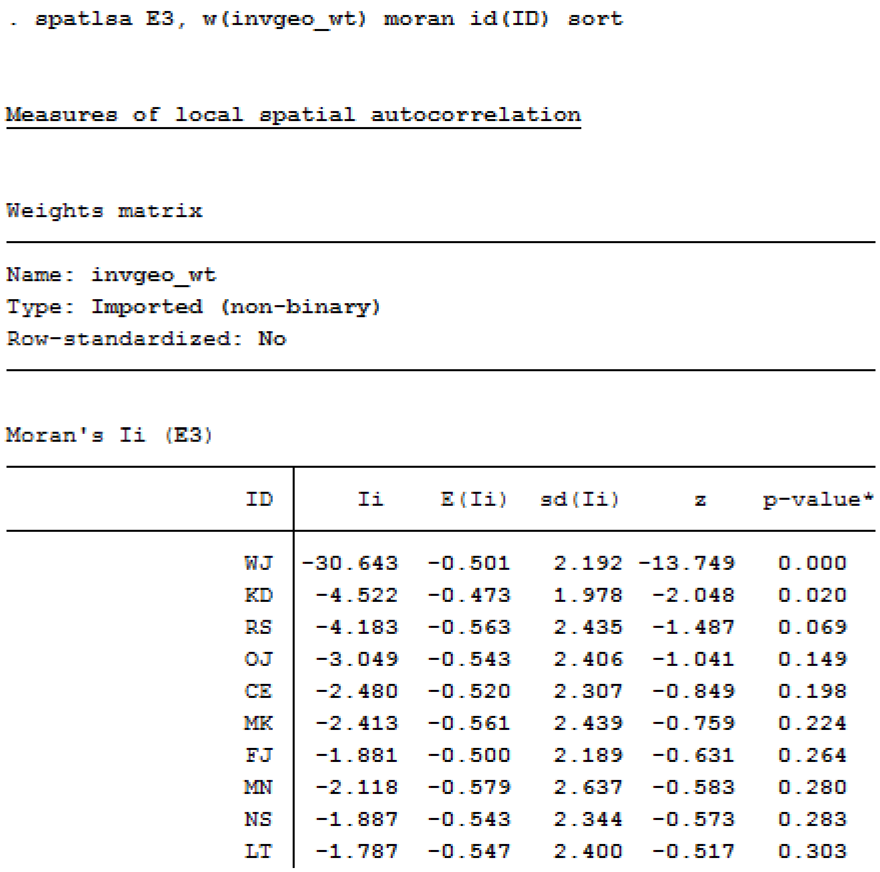


```
. use attributes2011.dta

. spatgsa E3, w(invgeo_wt) moran


Measures of global spatial autocorrelation


Weights matrix
--------------------------------------------------------------
Name: invgeo_wt
Type: Imported (non-binary)
Row-standardized: No
--------------------------------------------------------------

Moran's I
--------------------------------------------------------------
           Variables |      I      E(I)    sd(I)      z    p-value*
---------------------+----------------------------------------
                  E3 |  -0.024  -0.014    0.006  -1.585   0.056
--------------------------------------------------------------
*1-tail test
```

The Moran statistic is negative, indicating a positive global autocorrelation of student’s grades on the final exam. The value calculated in Stata (-0.024) is identical to the one calculated in UCINET to at least three decimal places (see Figure 5.14). The significance level reported by Stata, however, is slightly more conservative than the UCINET result. Minor differences are not unusual given the use of permutation trials.

Calculating the local version of the statistic can be done in the same working session with *spatlsa*, as shown in figure 5.21.

Figure 5.21. Stata Syntax and Output for Local Autocorrelation of Final Exam Using *spatlsa*



```
. spatlsa E3, w(invgeo_wt) moran id(ID) sort


Measures of local spatial autocorrelation


Weights matrix
------------------------------------------------------------
Name: invgeo_wt
Type: Imported (non-binary)
Row-standardized: No
------------------------------------------------------------

Moran's Ii (E3)
------------------------------------------------------------
                 ID |      Ii    E(Ii)   sd(Ii)        z   p-value*
------------------------------------------------------------
                 WJ | -30.643   -0.501    2.192  -13.749    0.000
                 KD |  -4.522   -0.473    1.978   -2.048    0.020
                 RS |  -4.183   -0.563    2.435   -1.487    0.069
                 OJ |  -3.049   -0.543    2.406   -1.041    0.149
                 CE |  -2.480   -0.520    2.307   -0.849    0.198
                 MK |  -2.413   -0.561    2.439   -0.759    0.224
                 FJ |  -1.881   -0.500    2.189   -0.631    0.264
                 MN |  -2.118   -0.579    2.637   -0.583    0.280
                 NS |  -1.887   -0.543    2.344   -0.573    0.283
                 LT |  -1.787   -0.547    2.400   -0.517    0.303
```

The syntax of the command is typical Stata: the procedure and variable list (multiple autocorrelations can be done in the same run) come first, followed by options after a comma. Here, the name of the weights matrix is supplied, the Moran statistic is specified (others are available, including Geary), the case name variable is identified to label the output, and a sort in order of decreasing positive autocorrelations is requested.

From the output, we see that the node WJ has an extremely strong positive autocorrelation of grades in his/her neighborhood, for example. The local autocorrelation of grades for each actor can be treated as an attribute of that actor describing something about the way in which they are embedded in the network. Indeed, one might think of the strength of the autocorrelation of grades as an indicator of the extent to which the actor has built ties with similar others, and/or been influenced by their neighbors. This is one way in which we might proceed to measure the degree of social influence on ego with regard to predicting ego's grade on the final exam. That is, ego's grade on the final exam might include a measure of how much ego is being influenced by his/her neighbor's exam grades.

5.4 Summary

In this chapter we've had a look at some simple methods for examining the association between dyadic and nodal variables (i.e. social relations or networks and attributes). Analyses like the ones discussed here cross "levels of analysis" and involve both the attributes of the two individual nodes involved in a relationship, as well as information about the structure of the network in which they are embedded.

In analyses that include both individual attributes and network relations, our interest may focus on either as the dependent variable.

Sometimes we are primarily interested in how the ways in which actors are embedded in networks provide constraints and opportunities that shape or *socially influence* their own attitudes and behaviors ("social influence" models, Chapter 6). Sometimes we are interested in how individual attributes affect the processes by which social structures are *selected* or built by making and breaking connections ("network selection" models, Chapters 7 and 8). In many cases, both kinds of processes are occurring simultaneously as networks and the individuals that they connect co-evolve (Chapter 9).

In the remaining chapters, we'll focus on modeling for these types of processes.

5.5 References

Getis, Arthur and J. K. Ord. 1992. "The analysis of spatial association by use of distance statistics." *Geographical Analysis*, 24, 189-206.

Griffith, Daniel. 1987. "Spatial Autocorrelation: A Primer". Washington, DC: Association of American Geographers Resource Publication.

McPherson, J. Miller and James R. Ranger-Moore. 1991. “Evolution on a Dancing Landscape: Organizations and Networks in Dynamic Blau Space.” *Social Forces*, 70(1): 19-42.

Pisati, Maurizio. 2001. “sg162: Tools for spatial data analysis.” *Stata Technical Bulletin,* 60: 21-37. In Stata Technical Bulletin Reprints, vol. 10, 277-298. College Station, TX: Stata Press.

Pisati, Maurizio. 2012. Spatial Data Analysis in Stata: An Overview. *2012 Italian Stata Users Group meeting*. Bologna, September 20-21, 2012. (slides available at: http://www.stata.com/meeting/italy12/abstracts/materials/it12_pisati.pdf)

Chapter 6. Network Influences on Attributes

6.1 Individual Attributes as Outcomes

A key insight of SNA is the seemingly obvious idea that an individual's attitudes and behaviors are affected by the attitudes and behaviors of those to whom the actor is connected. In order to understand or predict ego's attributes then, we need to take processes of social learning and social influence into account. The processes of learning, influence, and diffusion are all action *on* social networks.

In this chapter we'll look at some ways that social influence can be incorporated into statistical models that predict individuals' attributes and behaviors as outcomes. Models of this type include both the individual's own attributes, and the attributes of those to whom an individual is tied to as predictors. Models of how an individual's attributes may affect outcomes are commonplace, if not important. For example, women may be more likely to vote than men. What social network analysis adds to explanations of this type are the effects of social learning and influence. For example, while women may generally be more likely to vote than men, women who have more friends that are women might be even more likely to vote than those who have less gender homophilous networks.

The challenges of including information about ego's network in explanations of ego's behavior are more conceptual and theoretical than they are methodological. Models of individual outcomes for actors embedded in networks use individuals as units of analysis, and use conventional general linear modeling approaches. Data about how ego's network affects ego's behavior are included as attributes of ego – though they measure structural properties of the graph. For example, assuming that gender is predictive of voting behavior, when explaining whether a person votes or not, one might very well want to include information about the size of ego's friendship network as well as the proportion of women in their friendship network.

In many cases, the most important attribute of ego's network is whether the alters in the network display the behavior or attitude in question. If we want to explain variation in voting, we would probably hypothesize that as the number or proportion of ego's neighbors who vote increases, the likelihood that ego will vote increases. The technical term for this is "network autoregression." This is the tendency for a node to have scores on the dependent variable that are more similar to those of their network neighbors than to random others as a result of causal processes (social learning, influence, diffusion, etc.). Network autoregressive processes are modeled by including measures of the prevalence of the outcome in ego's neighborhood. Sometimes we may wish to include indirect, as well as direct influences. For more complex processes operating over broader social space, "spatial lag" models may be helpful.

There is also the possibility that a positive autocorrelation between ego and ego's close neighbors arises from local "error" or unobserved latent variables affecting both ego and alters. One may overstate the significance of autoregression (social learning or influence) in the presence of spatially autocorrelated error.

Below, we'll first take a look at some common approaches to incorporating information about how an individual is embedded in a network into models predicting an individual's attitudes or behavior. Many social learning and social influence processes that operate only locally to produce autoregression can be specified fairly easily. For influence processes that operate over somewhat broader social spaces, we may wish to borrow spatial autoregressive modeling approaches from econometrics and geostatistics. And, finally, we'll take a look at how spatially autocorrelated error or mis-specification can be controlled.

6.2 Preparing for Analysis

Setting up the data for analyses that predict the attributes of nodes is fairly straightforward. Most analyses represent independent variables as attributes of the individual, and use regular generalized linear modeling, with permutation test approaches to assess significance. We'll talk first about some typical kinds of predictor variables such as other attributes of the individual along with measures of how they are embedded in the network. Then we'll discuss spatial weights, which are necessary for social-spatial autoregressive and autocorrelated error models.

6.2.1 Dependent Variables: Levels of Measurement and Distributional Shapes

Models of network influence usually focus on some time-varying attribute of individuals, such as attitudes, behaviors, or conditions. We might be interested in examining the role of network influence on attitudes toward some issue. Do people who have many friends who believe that the earth is flat, also tend to believe it is flat? Sometimes behaviors are the focus. Are students who are friendly with other students who have poor academic performance more likely to perform badly themselves? And, sometimes we might focus on some kind of material condition. Are the social networks of people who own Lamborghini automobiles likely to contain others who also own these? Recently, and controversially, it has been suggested that happiness and depression may be transmitted by way of social networks.

The outcome in our models is an attribute of individuals, and can be scaled at any level of measurement. Depending on how the dependent variable is measured, appropriate versions of distributional family and link functions of generalized linear models can be used (except for spatial autocorrelated error and spatial autoregression models which have somewhat more limited software at this writing). In our examples, we'll focus on two measures of the academic success of our students. As a binary measure, we'll examine which factors may affect whether the student passed or failed the final exam (that is, earned a grade of 70% or more). As a continuous measure, we'll focus on the same final exam, but analyze the actual score earned.

6.2.2 Independent Variables: Attributes

Conventional (non-SNA) models of attitudes and behavior focus on individual level factors as predictors. For example, men and women may be expected to differ on an outcome, people of different ethnicities or religions may be hypothesized to vary, etc. Individual attributes are used in social influence analyses in exactly the same ways, and there is nothing much more to say about them here. Individual attributes are simply coded and entered as they might be in any other analysis (see Chapter 3).

In our student data that we introduced in Chapter 2, we don't have very much information about the individual attributes of the students. Only self-identified gender and ethnicity were recorded. We might predict that women students would out-perform men students on the final exam in our class, as they tend to do in most classes, due to differential socialization and selection processes that occurred prior to entering our course. Similarly, we might expect differences by ethnicity.

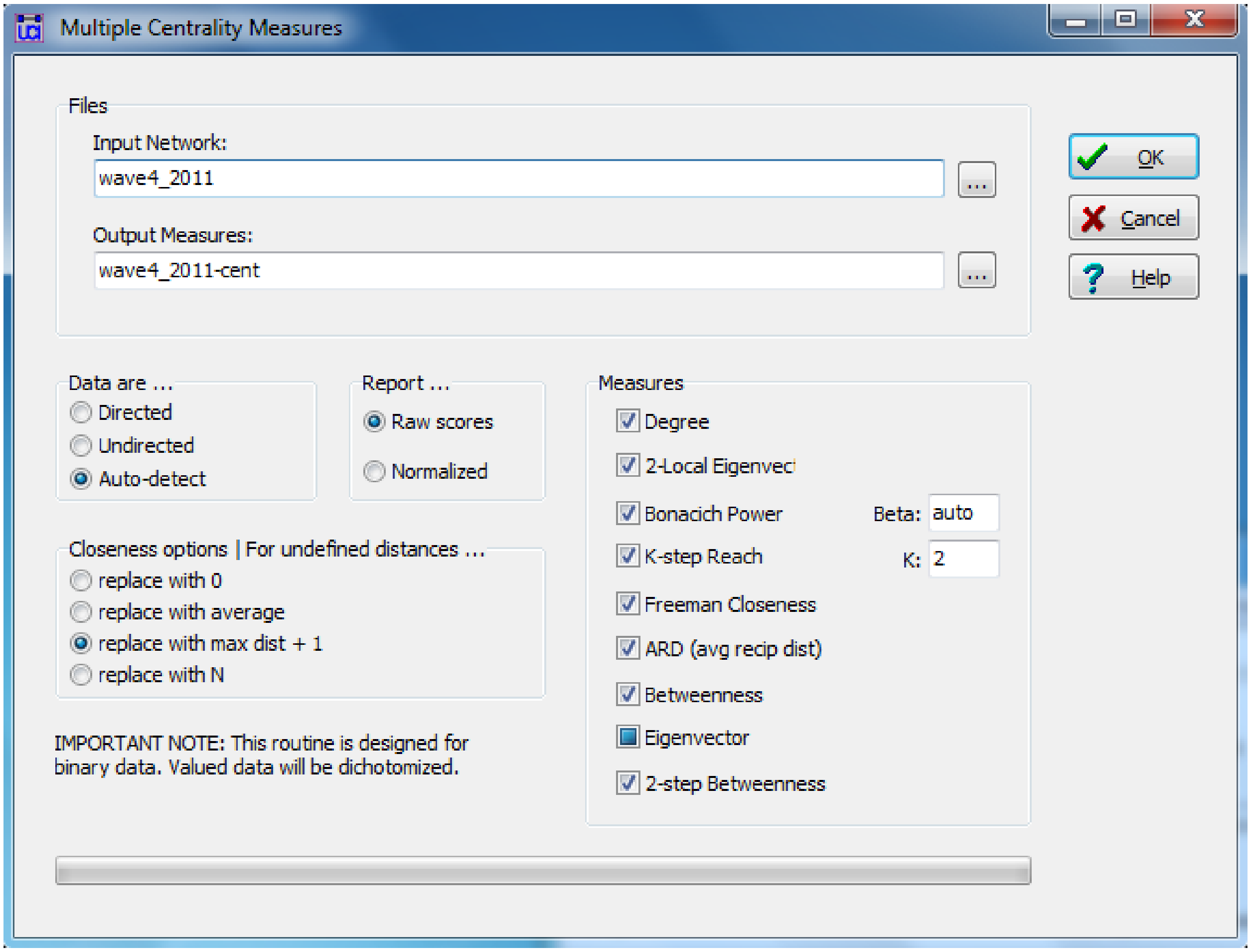
6.2.3 Independent Variables: Measures of Network Embedding

Conventional analyses of an individual's attitudes and behaviors rarely pay much attention to explicitly measuring variations in an individual's social networks. SNA, of course, focuses on describing and indexing variation in the ways that individuals are embedded in networks. People who have many friends may be easier to reach with a message than those who are less connected. Actors who are tied to others who are tied to one another may be more difficult to convert because of strong countervailing pressures. Actors who are embedded in networks that are highly homophilous may be more subject to conformity pressures. Actors who are key-players, or central and prominent figures in networks, may be more likely to have conventional attitudes and behaviors.

Social influence models may include a number of different measures about how an actor is connected as predictors of their attitudes and behaviors. Just like the attribute variables discussed above, these measures of network embedding are nodal variables describing the ways in which the nodes of a network are embedded. The authors of the UCINET software have provided tools that are useful for building variables that describe how nodes are embedded in the network. Taking a brief look at these tools suggests some common types of network-contextual variables that one might want to include in an analysis. The UCINET tools also have the advantage of calculating numerous network measures that are saved as files, and can be appended to our attribute file for use as predictors.

Network>Multiple Measures> Node Level produces a number of measures of how connected nodes are, their distance from other nodes in the network, and their centrality (defined in a variety of ways). Figure 6.1 shows the dialog.

Figure 6.1. UCINET Dialog for Node Level Measures of Network Embedding



Multiple Centrality Measures

Files

Input Network: wave4_2011

Output Measures: wave4_2011-cent

Data are …
○ Directed
○ Undirected
● Auto-detect

Report …
● Raw scores
○ Normalized

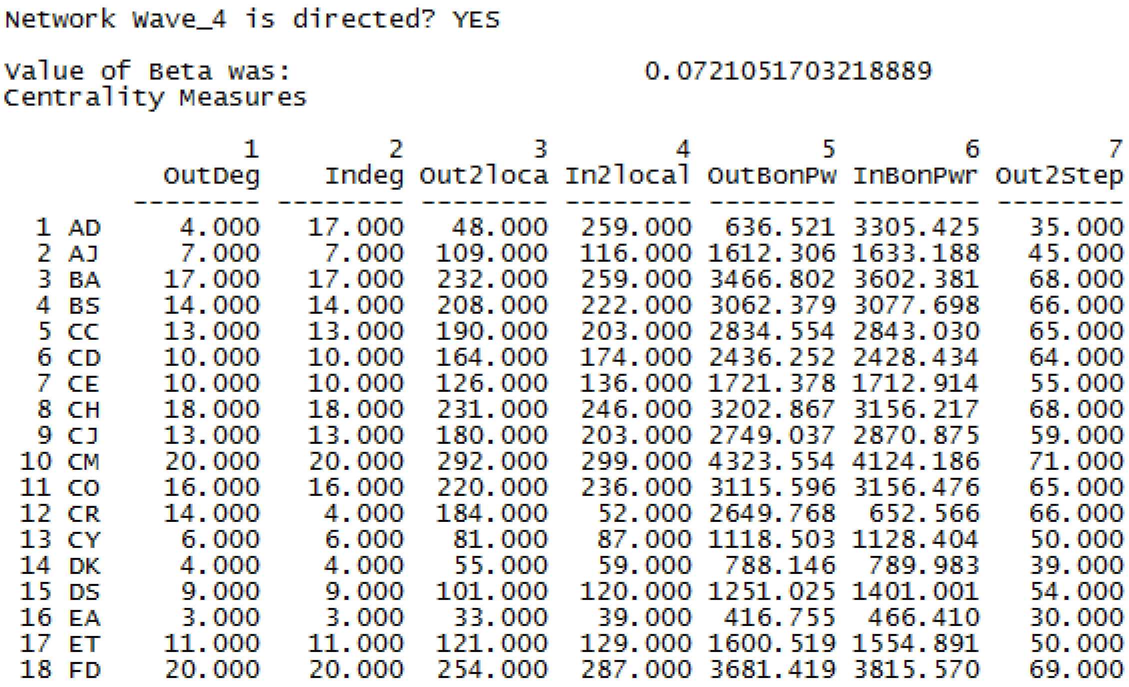
Closeness options | For undefined distances …
○ replace with 0
○ replace with average
● replace with max dist + 1
○ replace with N

IMPORTANT NOTE: This routine is designed for binary data. Valued data will be dichotomized.

Measures
☑ Degree
☑ 2-Local Eigenvec
☑ Bonacich Power Beta: auto
☑ K-step Reach K: 2
☑ Freeman Closeness
☑ ARD (avg recip dist)
☑ Betweenness
■ Eigenvector
☑ 2-step Betweenness

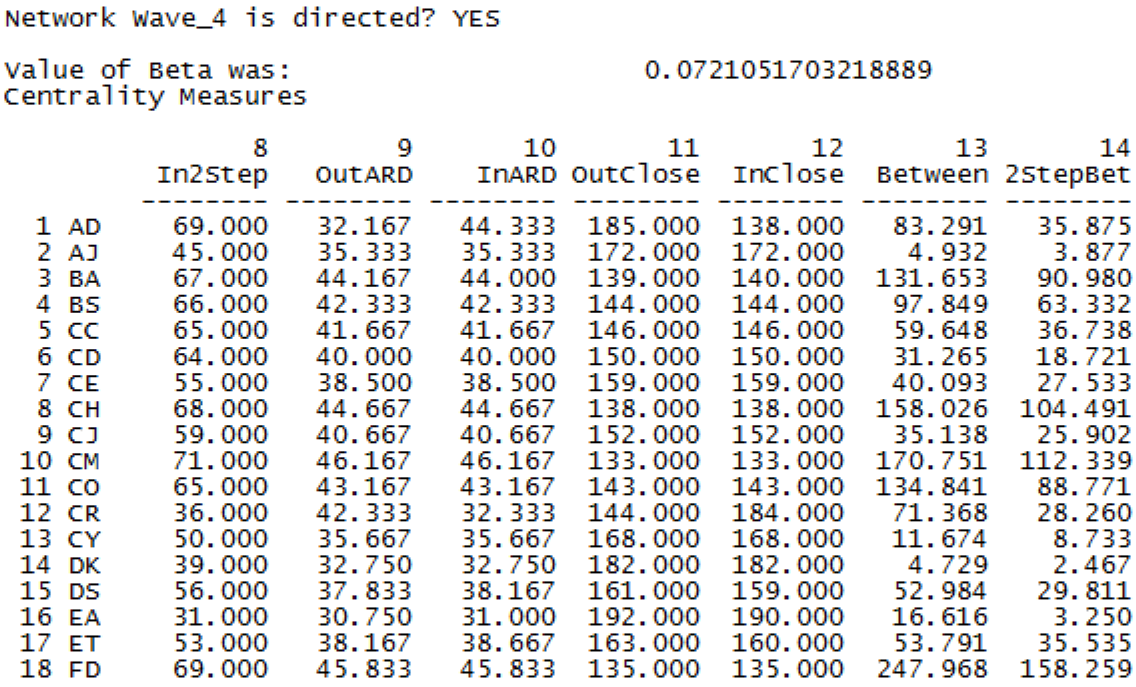
In the dialog, we're using the student data introduced in Chapter 2. We've specified that the input file is the acquaintanceship network at the time of the course final exam (wave 4). Since we will want to save the output and append it to the file of individual attributes, we've used the default file name for the output. We could specify the data as directed, or let the program detect that it is (auto-detect) given the fact that the acquaintanceship matrix is asymmetric across its diagonal. UCINET can calculate normalized scores, which may be useful for comparison across different networks. Since we are only concerned with the one class, we'll keep the original metrics (note "Raw scores" is selected). The last panel (lower right), lets us choose various measures of distance and centrality, which we will discuss with the output that is shown in Figures 6.2.1 and 6.2.2 (the output was too wide to fit into a single figure so it was split into two parts).

Figure 6.2.1. UCINET Output of Wave 4 Centralization Measures (Part 1)



```
Network Wave_4 is directed? YES

Value of Beta was:                 0.0721051703218889
Centrality Measures

                1        2        3        4        5        6        7
           OutDeg    Indeg Out2loca In2local OutBonPw InBonPwr Out2Step
         -------- -------- -------- -------- -------- -------- --------
 1 AD       4.000   17.000   48.000  259.000  636.521 3305.425   35.000
 2 AJ       7.000    7.000  109.000  116.000 1612.306 1633.188   45.000
 3 BA      17.000   17.000  232.000  259.000 3466.802 3602.381   68.000
 4 BS      14.000   14.000  208.000  222.000 3062.379 3077.698   66.000
 5 CC      13.000   13.000  190.000  203.000 2834.554 2843.030   65.000
 6 CD      10.000   10.000  164.000  174.000 2436.252 2428.434   64.000
 7 CE      10.000   10.000  126.000  136.000 1721.378 1712.914   55.000
 8 CH      18.000   18.000  231.000  246.000 3202.867 3156.217   68.000
 9 CJ      13.000   13.000  180.000  203.000 2749.037 2870.875   59.000
10 CM      20.000   20.000  292.000  299.000 4323.554 4124.186   71.000
11 CO      16.000   16.000  220.000  236.000 3115.596 3156.476   65.000
12 CR      14.000    4.000  184.000   52.000 2649.768  652.566   66.000
13 CY       6.000    6.000   81.000   87.000 1118.503 1128.404   50.000
14 DK       4.000    4.000   55.000   59.000  788.146  789.983   39.000
15 DS       9.000    9.000  101.000  120.000 1251.025 1401.001   54.000
16 EA       3.000    3.000   33.000   39.000  416.755  466.410   30.000
17 ET      11.000   11.000  121.000  129.000 1600.519 1554.891   50.000
18 FD      20.000   20.000  254.000  287.000 3681.419 3815.570   69.000
```

Figure 6.2.2. UCINET Output of Wave 4 Centralization Measures (Part 2)



```
Network Wave_4 is directed? YES

Value of Beta was:                 0.0721051703218889
Centrality Measures

                8        9       10       11       12       13       14
          In2Step   OutARD    InARD OutClose  InClose  Between 2StepBet
         -------- -------- -------- -------- -------- -------- --------
 1 AD      69.000   32.167   44.333  185.000  138.000   83.291   35.875
 2 AJ      45.000   35.333   35.333  172.000  172.000    4.932    3.877
 3 BA      67.000   44.167   44.000  139.000  140.000  131.653   90.980
 4 BS      66.000   42.333   42.333  144.000  144.000   97.849   63.332
 5 CC      65.000   41.667   41.667  146.000  146.000   59.648   36.738
 6 CD      64.000   40.000   40.000  150.000  150.000   31.265   18.721
 7 CE      55.000   38.500   38.500  159.000  159.000   40.093   27.533
 8 CH      68.000   44.667   44.667  138.000  138.000  158.026  104.491
 9 CJ      59.000   40.667   40.667  152.000  152.000   35.138   25.902
10 CM      71.000   46.167   46.167  133.000  133.000  170.751  112.339
11 CO      65.000   43.167   43.167  143.000  143.000  134.841   88.771
12 CR      36.000   42.333   32.333  144.000  184.000   71.368   28.260
13 CY      50.000   35.667   35.667  168.000  168.000   11.674    8.733
14 DK      39.000   32.750   32.750  182.000  182.000    4.729    2.467
15 DS      56.000   37.833   38.167  161.000  159.000   52.984   29.811
16 EA      31.000   30.750   31.000  192.000  190.000   16.616    3.250
17 ET      53.000   38.167   38.667  163.000  160.000   53.791   35.535
18 FD      69.000   45.833   45.833  135.000  135.000  247.968  158.259
```

The number of ties actors have to others in the network, and how close they are to others, could affect whether they adopt an attitude or behavior. From the output in Figure 6.2.1, we see that AJ has fewer connections than actor BA, both from others naming them as acquaintances (in degree; column 1) and others that they've named (out degree; column 2). UCINET also provides measures of out degree and in degree of an actor's direct acquaintances. In column 3, Out2local tells us the total number of acquaintances named by each acquaintance named by ego. For example, because the out degree of AD is equal to four, we know AD named four others as acquaintances (whether they named AD or not). The total number of others named by those four is captured by Out2local and is equal to 48. Similarly, In2local (column 4) tells us the number of acquaintances that named the acquaintances of ego. So because 17 students named AD as an acquaintance (in degree = 17), we know those 17 were named as acquaintances 259 times from In2local.

The local two-degree measures just discussed seem a little odd. We know there are only 75 students total in the data set, yet the 17 in degree acquaintances of AD were named 259 times? This is due to the potential for repetition in naming others (e.g. multiple students might name all 17 students that name AD, generating a large value of In2local). If instead we are interested in the unique number of students that are separated from ego by two steps ("friends of my friends"), we can look at Out2step (column 7 in Fig. 6.2.1) and In2step (column 8 in Fig. 6.2.2). For example, though the 17 students that name AD are named 259 times, there are only 69 unique students that name them (In2step). So AD is connected to 69 of the 75 students in the class by only two steps of separation, at least for inward ties (pointing toward AD). AD is only connected to 35 students via two steps of outward ties (Out2step).

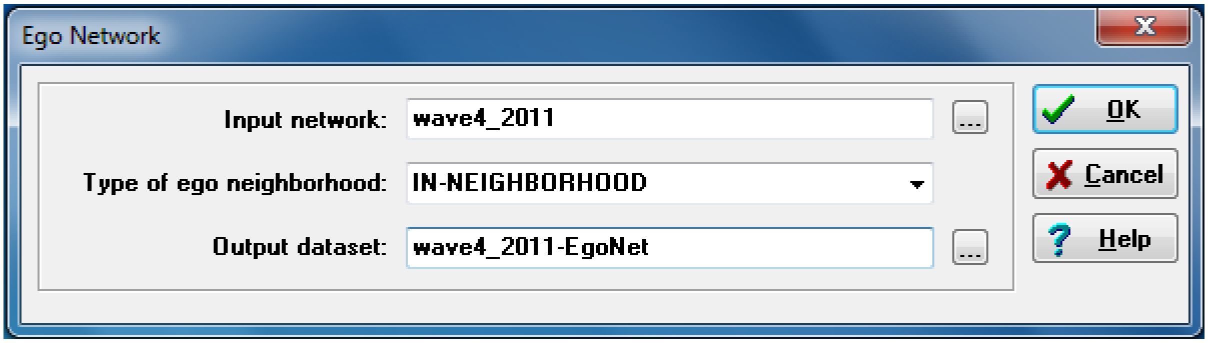
The average distance from ego to all others in the network that are reachable (OutARD, column 9 in Fig. 6.2.2) and the average distance to ego from all others in the network that can reach ego (InARD, column 10 in Fig. 6.2.2) are provided. Also, included are Outclose (column 11 in Fig. 6.2.2) and InClose (column 12 in Fig. 6.2.2). OutClose tells us the out-closeness which can be thought of as the degree to which ego can reach all other nodes in the network via short path lengths. In-closeness, then, is the degree to which ego can be reached by all other nodes via short path lengths. All of these measures are possible ways of thinking about the density or closeness of each actor to the broader network. Social influence models often hypothesize that actors who are more connected, in one way or another, are more likely to have attitudes and behaviors that are closer to the mean or mode.

The output also shows measures of ego's centrality in the network. The Bonacich power measures (as well as a similar measure that's not shown, the eigenvector centrality) indicate whether ego is connected to other well-connected actors. If ego has high in-centrality (e.g. column 6 in Fig. 6.2.1), i.e. they are being influenced by other influential actors, they may be "constrained." If ego has high out-centrality (column 5 in Fig. 6.2.1), they are able to exert influence on other influential actors. Actors who have high "betweenness" centrality (columns 13 and 14 in Fig 6.2.2) are acting as transmitters or brokers of ties between other pairs of actors. Being a "broker" may be a source of power, and hence make an actor less constrained and more influential.

For many social processes, the way that an actor is embedded in their local neighborhood of the social network can be more important than their overall centrality or distance from others. UCINET has a number of tools for describing ego-networks (the actors that are connected to ego at a short distance, and the connections among them). Metrics on each actor's ego-network can provide some interesting insights into variation in ego's attitudes and behaviors.

Network>Ego Networks>Egonet Basic Measures in UCINET allows varying definitions of neighborhoods, and calculates a number of measures describing the topology of each node's local network. Figure 6.3 shows a dialog for the acquaintanceship data at the time of the final exam.

Figure 6.3. UCINET Dialog of Basic Egonet Metrics for Wave 4 Acquaintanceship



Ego Network

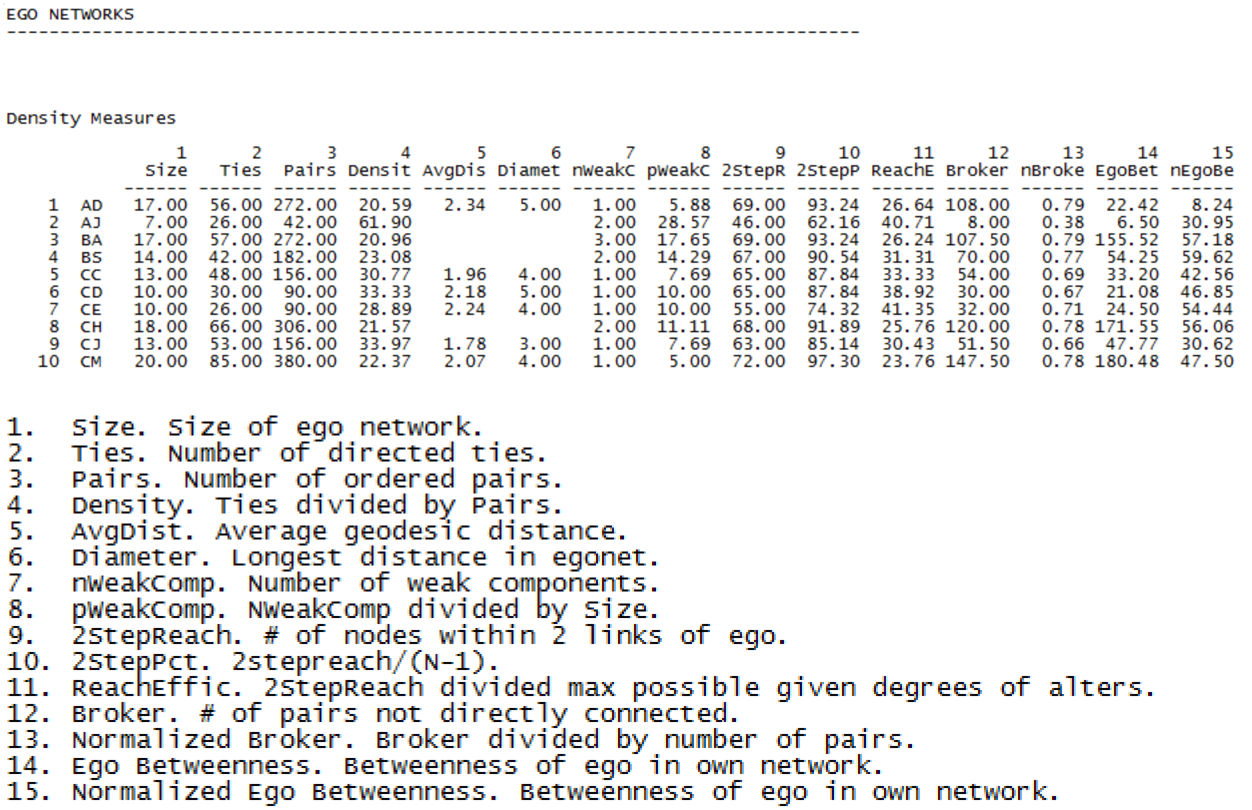
Input network: wave4_2011

Type of ego neighborhood: IN-NEIGHBORHOOD

Output dataset: wave4_2011-EgoNet

In the dialog we've identified the acquaintanceship matrix at wave 4 (the time of the final exam). Since the network is directed, we can choose to define it by relationships directed at ego (the in-neighborhood), relationships from ego to alters (out-neighborhood), or to include both. We've chosen to define ego's neighborhood as any actors that claim ego as an acquaintance (and any ties among these alters). The in-neighborhood is a reasonable choice because we are focusing on how ego is influenced by alters (rather than how influential ego is with alters). The output dataset was named and saved to be used later, so that it can be appended to the attributes dataset to act as a set of independent variables for predicting ego's attributes. Figure 6.4 shows the output for the first 10 of 75 egos.

Figure 6.4. UCINET Output of Basic Egonet Metrics for Wave 4 Acquaintanceship



```
EGO NETWORKS
------------------------------------------------------------------------------


Density Measures

                 1      2      3      4      5      6      7      8      9     10     11     12     13     14     15
              Size   Ties  Pairs Densit AvgDis Diamet nWeakC pWeakC 2StepR 2StepP ReachE Broker nBroke EgoBet nEgoBe
            ------ ------ ------ ------ ------ ------ ------ ------ ------ ------ ------ ------ ------ ------ ------
  1   AD     17.00  56.00 272.00  20.59   2.34   5.00   1.00   5.88  69.00  93.24  26.64 108.00   0.79  22.42   8.24
  2   AJ      7.00  26.00  42.00  61.90                 2.00  28.57  46.00  62.16  40.71   8.00   0.38   6.50  30.95
  3   BA     17.00  57.00 272.00  20.96                 3.00  17.65  69.00  93.24  26.24 107.50   0.79 155.52  57.18
  4   BS     14.00  42.00 182.00  23.08                 2.00  14.29  67.00  90.54  31.31  70.00   0.77  54.25  59.62
  5   CC     13.00  48.00 156.00  30.77   1.96   4.00   1.00   7.69  65.00  87.84  33.33  54.00   0.69  33.20  42.56
  6   CD     10.00  30.00  90.00  33.33   2.18   5.00   1.00  10.00  65.00  87.84  38.92  30.00   0.67  21.08  46.85
  7   CE     10.00  26.00  90.00  28.89   2.24   4.00   1.00  10.00  55.00  74.32  41.35  32.00   0.71  24.50  54.44
  8   CH     18.00  66.00 306.00  21.57                 2.00  11.11  68.00  91.89  25.76 120.00   0.78 171.55  56.06
  9   CJ     13.00  53.00 156.00  33.97   1.78   3.00   1.00   7.69  63.00  85.14  30.43  51.50   0.66  47.77  30.62
 10   CM     20.00  85.00 380.00  22.37   2.07   4.00   1.00   5.00  72.00  97.30  23.76 147.50   0.78 180.48  47.50

1.   Size. Size of ego network.
2.   Ties. Number of directed ties.
3.   Pairs. Number of ordered pairs.
4.   Density. Ties divided by Pairs.
5.   AvgDist. Average geodesic distance.
6.   Diameter. Longest distance in egonet.
7.   nWeakComp. Number of weak components.
8.   pWeakComp. NWeakComp divided by Size.
9.   2StepReach. # of nodes within 2 links of ego.
10.  2StepPct. 2stepreach/(N-1).
11.  ReachEffic. 2StepReach divided max possible given degrees of alters.
12.  Broker. # of pairs not directly connected.
13.  Normalized Broker. Broker divided by number of pairs.
14.  Ego Betweenness. Betweenness of ego in own network.
15.  Normalized Ego Betweenness. Betweenness of ego in own network.
```

Most of the common metrics for ego-networks (egonets, for short) get at the same ideas as the global measures we saw above. Student AJ, for example, has 7 students who name him/her as an acquaintance (column 1). Among those 7 others (not including AJ), there are 42 possible ties (column 3), of which 26 are actually present (column 2). The density of ego's local neighborhood, then, is 61.9% (column 4).

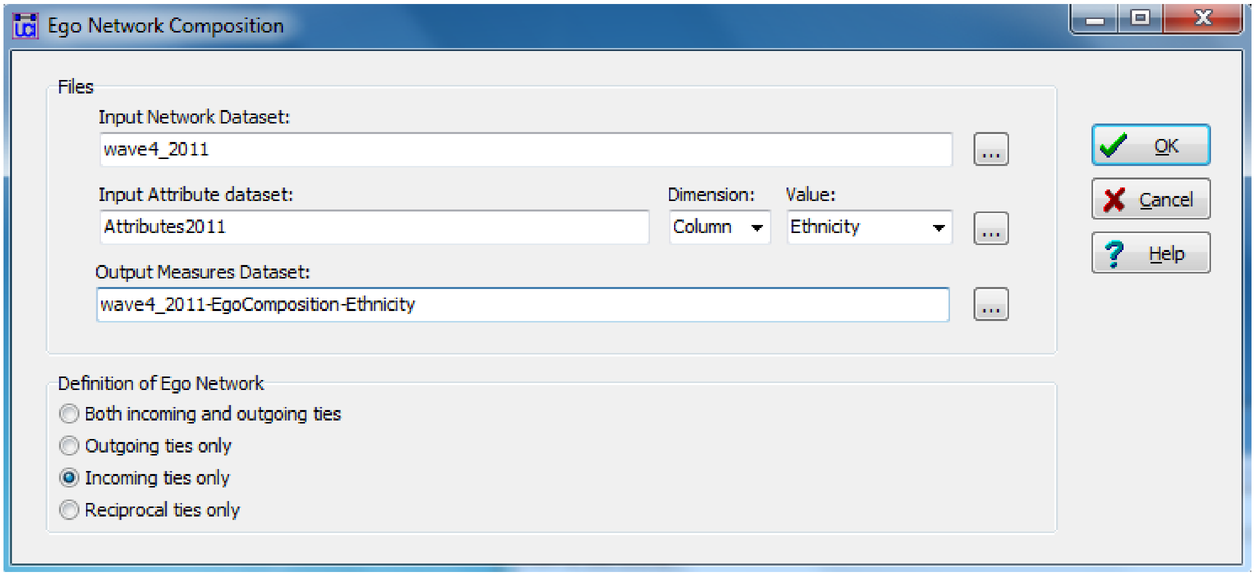
The density of an egonet, leaving ego out of it, is also called the "clustering coefficient," and it is often useful as a measure of the "open-ness" of ego's network. If ego is known by others who are connected to one another, we may observe "clique" like pressures for conformity and resistance to outside influences. If ego's network has low clustering, or is an "open" network, ego may be exposed to more diverse and competing pressures. The "granularity" of ego's local network, or the extent to which it is composed of groups of connected others, is also commonly indexed by the ratio of the number of weak components (groups of alters, all of whom are connected to one another; column 7). How close ego's network is to the network as a whole can be seen by looking at the percentage of all nodes that are within two steps of ego (the first step is in ego's local network, the second step reaches beyond; column 9). The "reach efficiency" measure (column 11) also addresses the idea of the open-ness of ego's network. The efficiency measure tells us the amount of contact ego gets with the larger network per member of ego's personal network. Low ratios indicate more network closure.

The last four measures deal with ego's influence or power in their local neighborhood. It is assumed for egonets that ego has more influence and greater autonomy to the extent that he/she controls the access of other members of the network to one another. Brokerage (column 12) and betweenness (column 14) are alternative ways of indexing the extent to which alters in ego's neighborhood are dependent on ego. Some additional measures of ego's autonomy and influence derived from Ronald Burt's work (1992) are available under *Network>Ego Networks> Structural Holes*. Some measures of the role that ego plays in connecting groups of actors with different attributes (e.g. acting as a "gate keeper" in relations between White and Asian students), developed by Gould and Fernandez (1989), are available under *Network> Ego Network> G+F Brokerage roles*.

The measures of ego's local network that we've looked at so far refer to its structure and ego's embedding in it. It may also be very important to understand the composition of ego's network (e.g. is ego tied to mostly women, or mostly to men? And, to what extent is ego tied to alters who are similar or dissimilar to ego, i.e. homophily?). UCINET has some helpful tools for constructing measures of network composition and homophily of the local context of each actor. These indexes can also be saved and appended to the attributes dataset.

Figure 6.5 shows the dialog for *Network>Ego Network>Egonet Composition>Categorical Alter Attributes* which can be used to build measures of the attributes of the others in ego's one-step neighborhood. There are two versions of this tool. One works with categorical alter attributes (like the gender or racial compositions of those tied to ego); the other with continuous alter attributes (like the average attendance rates or test scores of those tied to ego).

Figure 6.5. UCINET Dialog for Generating Egonet Composition Measures (Categorical Alter Attributes)



Input Network Dataset: wave4_2011

Input Attribute dataset: Attributes2011

Dimension: Column

Value: Ethnicity

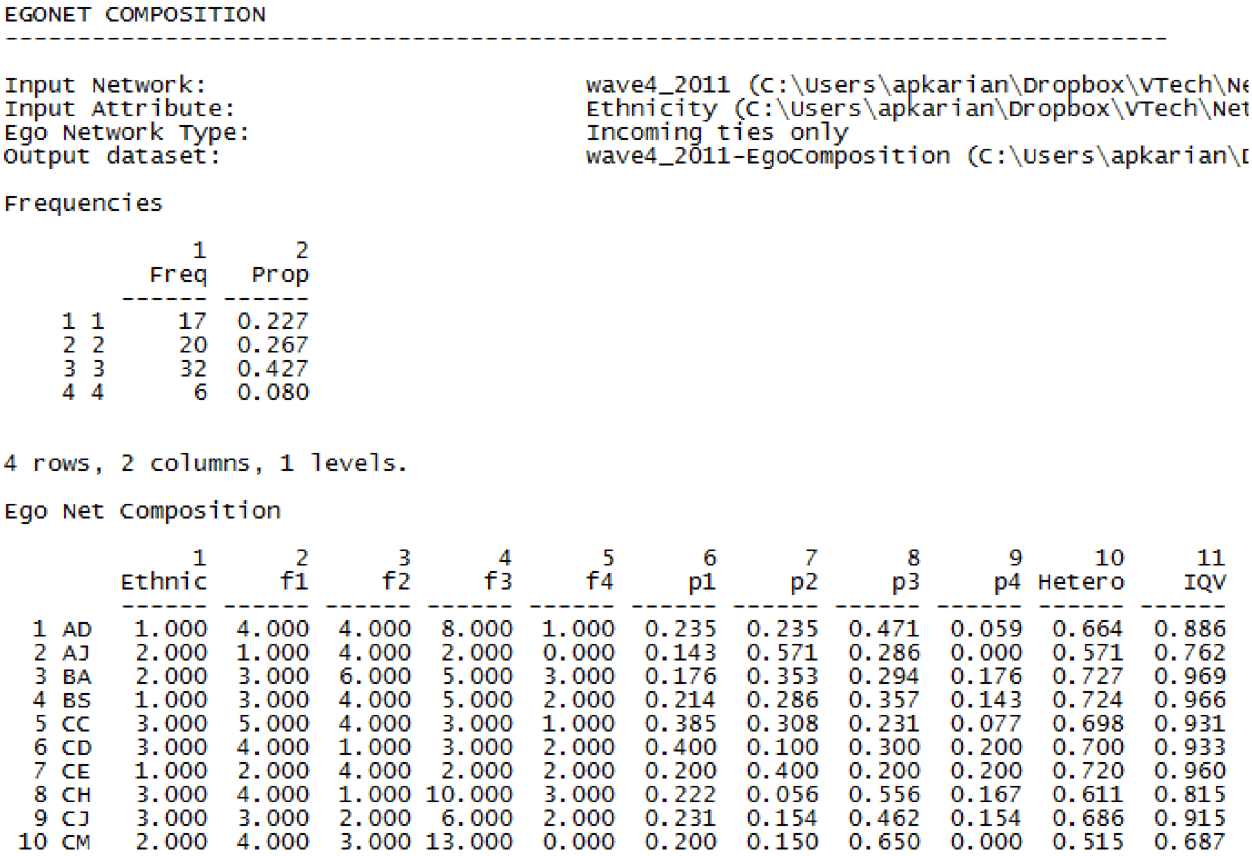
Output Measures Dataset: wave4_2011-EgoComposition-Ethnicity

Definition of Ego Network

- ○ Both incoming and outgoing ties
- ○ Outgoing ties only
- ● Incoming ties only
- ○ Reciprocal ties only

In this dialog, we focus on the acquaintanceship network at the time of the final examination and identify the attributes dataset and column that contains ego's ethnic identity. If we are interested in the social composition of actors who are influencing (rather than being influenced by) each student, we can select "incoming ties only." Figure 6.6 shows a portion of the results which are saved as a file that can be appended to the attribute file. We will save the output measures to be used later.

Figure 6.6. UCINET Output of Egonet Composition Measures (Categorical Alter Attributes)

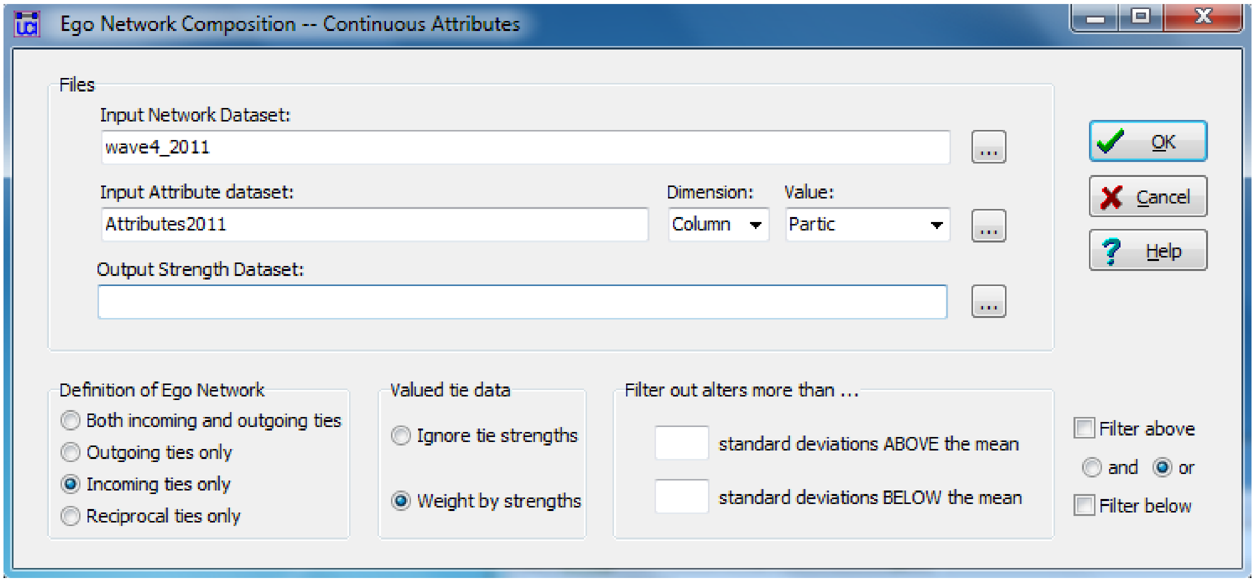


```
EGONET COMPOSITION
--------------------------------------------------------------------------------

Input Network:                          wave4_2011 (C:\Users\apkarian\Dropbox\VTech\Ne
Input Attribute:                        Ethnicity (C:\Users\apkarian\Dropbox\VTech\Net
Ego Network Type:                       Incoming ties only
Output dataset:                         wave4_2011-EgoComposition (C:\Users\apkarian\D

Frequencies

              1      2
           Freq   Prop
         ------ ------
    1 1      17  0.227
    2 2      20  0.267
    3 3      32  0.427
    4 4       6  0.080


4 rows, 2 columns, 1 levels.

Ego Net Composition

              1      2      3      4      5      6      7      8      9     10     11
         Ethnic     f1     f2     f3     f4     p1     p2     p3     p4 Hetero    IQV
         ------ ------ ------ ------ ------ ------ ------ ------ ------ ------ ------
  1 AD    1.000  4.000  4.000  8.000  1.000  0.235  0.235  0.471  0.059  0.664  0.886
  2 AJ    2.000  1.000  4.000  2.000  0.000  0.143  0.571  0.286  0.000  0.571  0.762
  3 BA    2.000  3.000  6.000  5.000  3.000  0.176  0.353  0.294  0.176  0.727  0.969
  4 BS    1.000  3.000  4.000  5.000  2.000  0.214  0.286  0.357  0.143  0.724  0.966
  5 CC    3.000  5.000  4.000  3.000  1.000  0.385  0.308  0.231  0.077  0.698  0.931
  6 CD    3.000  4.000  1.000  3.000  2.000  0.400  0.100  0.300  0.200  0.700  0.933
  7 CE    1.000  2.000  4.000  2.000  2.000  0.200  0.400  0.200  0.200  0.720  0.960
  8 CH    3.000  4.000  1.000 10.000  3.000  0.222  0.056  0.556  0.167  0.611  0.815
  9 CJ    3.000  3.000  2.000  6.000  2.000  0.231  0.154  0.462  0.154  0.686  0.915
 10 CM    2.000  4.000  3.000 13.000  0.000  0.200  0.150  0.650  0.000  0.515  0.687
```

In the above output, ethnicity = 1 refers to Whites, ethnicity = 2 refers to Hispanics, ethnicity = 3 refers to Asians, and ethnicity = 4 refers to African Americans. We see from the frequency table that the network as a whole is composed of 17 Whites, 20 Hispanics, 32 Asians, and 6 African Americans. From the first column, we can see that actor AJ identifies as Hispanic. AJ is identified as an acquaintance by one White student, four Hispanic students, two Asian students, and zero African American students (columns 2 through 5). The next four columns display the proportion of total incoming ties from students of each ethnic identity. One summary measure of heterogeneity (column 10 in the output) was developed by Peter Blau, and is described in UCINET as "1 minus the sum of the squares of the proportions of each value of the categorical variable in ego's network. For example, a person connected to equal numbers of men and women will have a Heterogeneity measure of 0.5, calculated as 1 - ((1/2)^2 + (1/2)^2))." This value can be thought of as a non-normalized index of qualitative variation (the interpretation varies by the number of categories of the discrete variable being measured). The last column provides the index of qualitative variation (column 11). Both measures suggest moderate heterogeneity in the ethnic identity of students who influence AJ.

Figure 6.7 shows the dialog for indexing the composition of ego's neighborhood on a continuous attribute. In this case, the rate of participation in the term paper project as assessed by members of the student's work group.

Figure 6.7. UCINET Dialog for Generating Egonet Composition Measures (Continuous Alter Attributes)



Ego Network Composition -- Continuous Attributes

Files

Input Network Dataset: wave4_2011

Input Attribute dataset: Attributes2011

Dimension: Column

Value: Partic

Output Strength Dataset:

Definition of Ego Network
- ○ Both incoming and outgoing ties
- ○ Outgoing ties only
- ● Incoming ties only
- ○ Reciprocal ties only

Valued tie data
- ○ Ignore tie strengths
- ● Weight by strengths

Filter out alters more than ...
- standard deviations ABOVE the mean
- standard deviations BELOW the mean

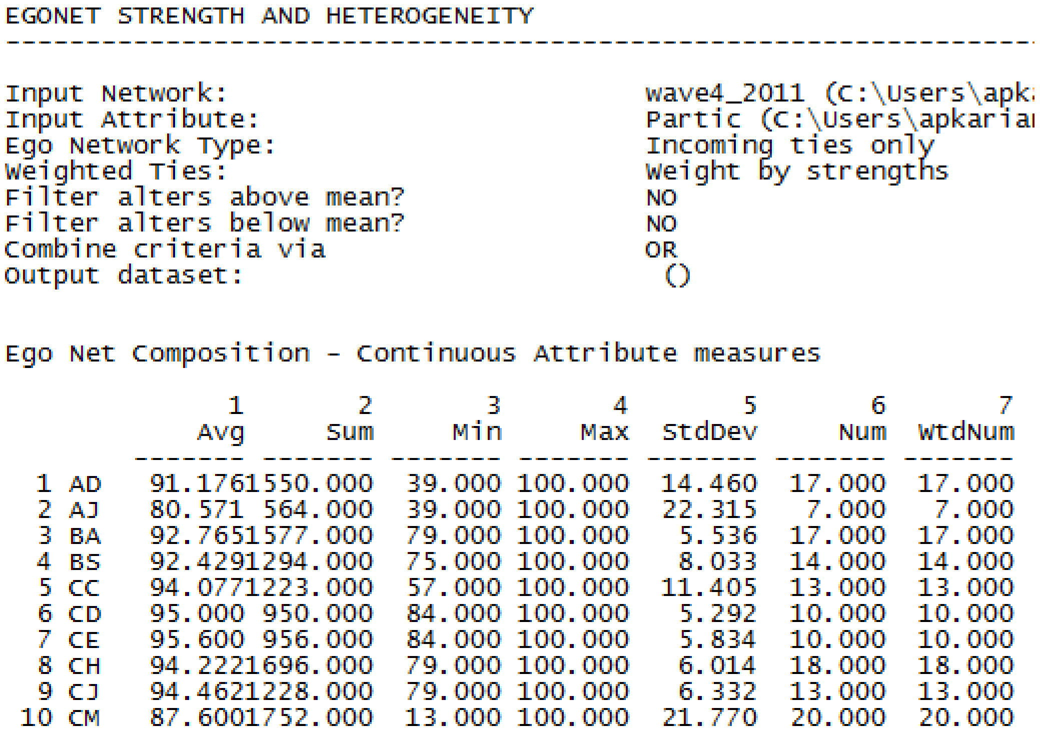
□ Filter above

○ and ● or

□ Filter below

As usual, we identify the network from which we want to extract the egonets (acquaintanceship at wave 4), and the file and column containing the student's attribute of participation grade in the final project. The output might be saved for appending to the attribute file. We've selected "Incoming ties only" to focus on the attributes of alters who are influencing ego. If the network was a valued network (e.g. "on a scale of 1 to 10, how well do you know X?"), we could weight the attributes of the alters proportional to tie strength (our network, however, is just a binary zero-one). Since the output will be means and standard deviations, alters with extreme scores could skew the results. Options are available to trim means and standard deviations. The output is shown in figure 6.8.

Figure 6.8. UCINET Output of Egonet Composition Measures (Continuous Alter Attributes)



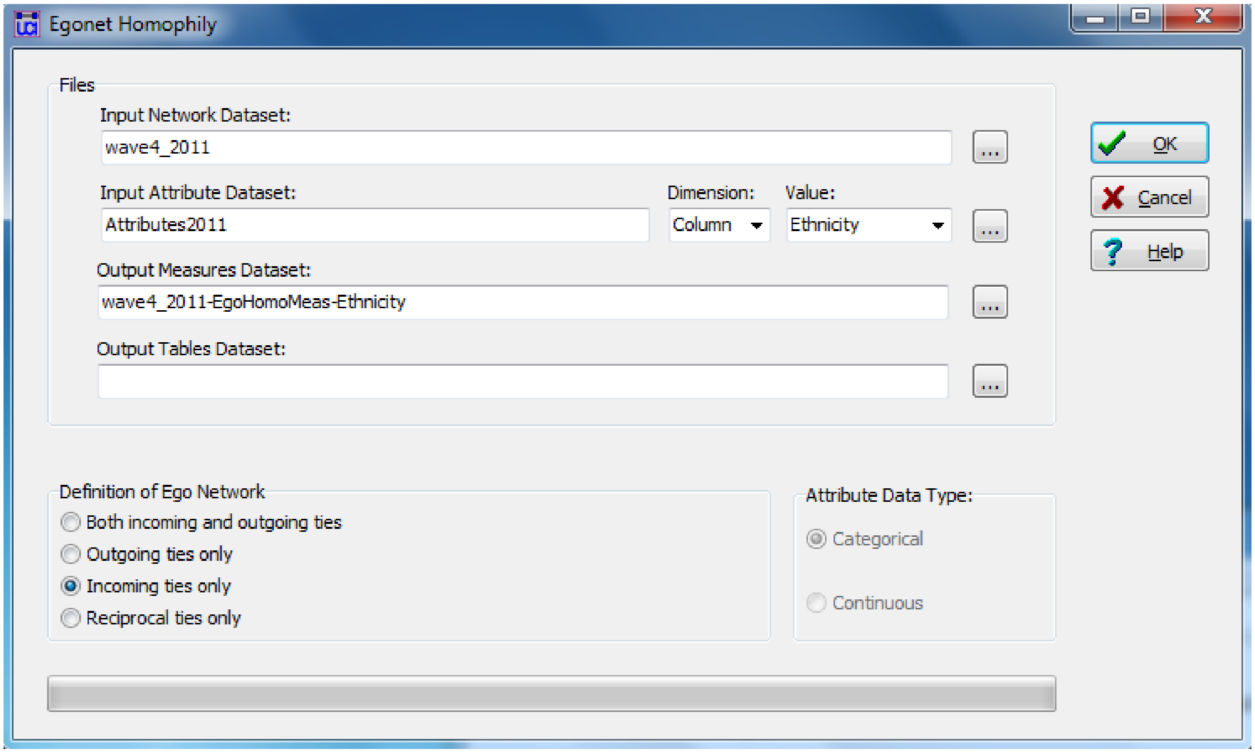
```
EGONET STRENGTH AND HETEROGENEITY
----------------------------------------------------------------

Input Network:                       wave4_2011 (C:\Users\apk
Input Attribute:                     Partic (C:\Users\apkaria
Ego Network Type:                    Incoming ties only
Weighted Ties:                       Weight by strengths
Filter alters above mean?            NO
Filter alters below mean?            NO
Combine criteria via                 OR
Output dataset:                        ()


Ego Net Composition - Continuous Attribute measures

                1       2       3       4       5       6       7
              Avg     Sum     Min     Max  StdDev     Num  WtdNum
          ------- ------- ------- ------- ------- ------- -------
 1 AD     91.1761550.000  39.000 100.000  14.460  17.000  17.000
 2 AJ     80.571  564.000  39.000 100.000  22.315   7.000   7.000
 3 BA     92.7651577.000  79.000 100.000   5.536  17.000  17.000
 4 BS     92.4291294.000  75.000 100.000   8.033  14.000  14.000
 5 CC     94.0771223.000  57.000 100.000  11.405  13.000  13.000
 6 CD     95.000  950.000  84.000 100.000   5.292  10.000  10.000
 7 CE     95.600  956.000  84.000 100.000   5.834  10.000  10.000
 8 CH     94.2221696.000  79.000 100.000   6.014  18.000  18.000
 9 CJ     94.4621228.000  79.000 100.000   6.332  13.000  13.000
10 CM     87.6001752.000  13.000 100.000  21.770  20.000  20.000
```

The first two columns appear to bleed into one another. Keep in mind that this command rounds to three decimal places, so the first column for AD reads 91.176 and the second column reads 1550.000. We can see from the output that AJ was identified as an acquaintance by seven others (column 6), all of whom were rated as not very active participants in the term project (average of 80.6 out of 100 possible; column 1). Their scores were also very diverse ranging from 39 to 100 with a standard deviation of 22.3. Actor CD, on the other hand, was being influenced by 10 alters who had a much higher average participation rate (95.0), who were also much less diverse in this regard (ranging from 84 to 100, with a standard deviation of only 5.3).

The ego-net composition tools are potentially quite important in the analysis of social influence, as they enable us to index the attributes of those who are directly connected to ego in the network. We may be interested in both the central tendency (what kind of alter is typical for ego? what is the mean or average score of the alters of ego?), and in diversity or variation among the alters.

Building on this idea of homogeneity or diversity of the attributes of those influencing alter, SNA particularly focuses attention on “homophily” or the extent to which ego is similar to alters. Students who are tied to other students who are very similar to themselves may be more subject to stronger constraints but may also have higher levels of social support. UCINET, again, provides a fairly convenient tool for indexing the homophily of ego’s network with *Network>Ego Networks>Egonet Homophily*. A dialog for this tool is shown in figure 6.9.

Figure 6.9. UCINET Dialog for Generating Egonet Homophily Measures



Egonet Homophily

Files

Input Network Dataset: wave4_2011

Input Attribute Dataset: Attributes2011 Dimension: Column Value: Ethnicity

Output Measures Dataset: wave4_2011-EgoHomoMeas-Ethnicity

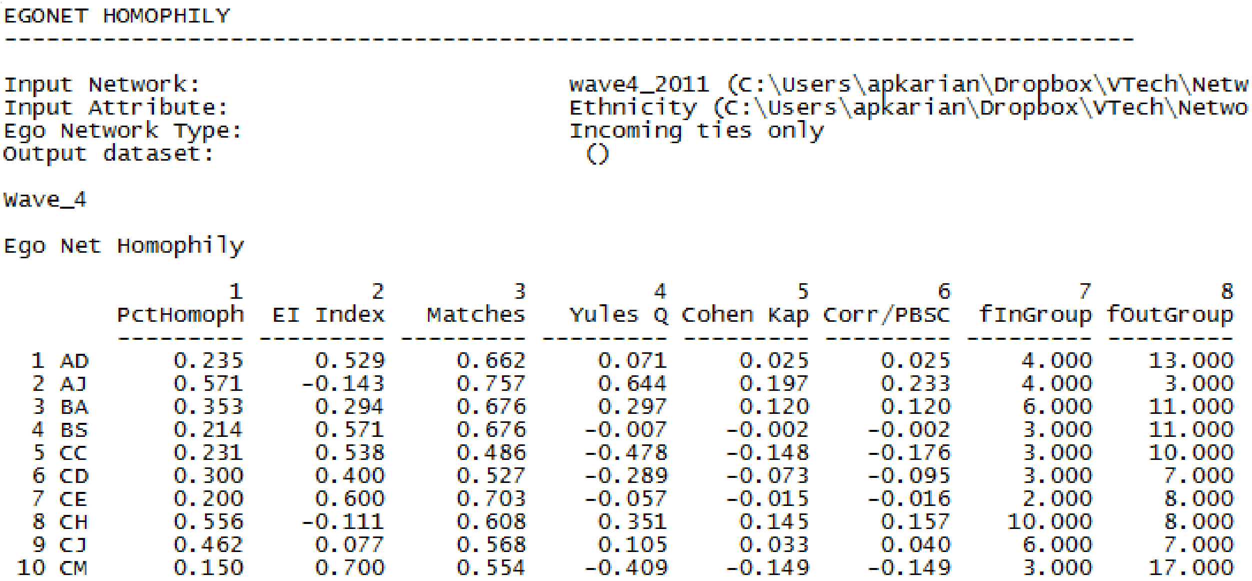
Output Tables Dataset:

Definition of Ego Network
○ Both incoming and outgoing ties
○ Outgoing ties only
● Incoming ties only
○ Reciprocal ties only

Attribute Data Type:
● Categorical
○ Continuous

In this dialog, we’ve asked for measures of the similarity of the ethnic composition of ego’s network to ego’s own ethnic identity, using the acquaintanceship network at wave 4 to identify incoming tie egonets. As always, the results shown in figure 6.10 can be saved to a file, and then appended to ego’s attributes to be used later.

Figure 6.10. UCINET Output of Egonet Homophily Measures



```
EGONET HOMOPHILY
--------------------------------------------------------------------------------

Input Network:                          wave4_2011 (C:\Users\apkarian\Dropbox\VTech\Netw
Input Attribute:                        Ethnicity (C:\Users\apkarian\Dropbox\VTech\Netwo
Ego Network Type:                       Incoming ties only
Output dataset:                          ()

Wave_4

Ego Net Homophily

                  1         2         3         4         5         6         7         8
          PctHomoph  EI Index   Matches   Yules Q Cohen Kap Corr/PBSC  fInGroup fOutGroup
          --------- --------- --------- --------- --------- --------- --------- ---------
 1 AD         0.235     0.529     0.662     0.071     0.025     0.025     4.000    13.000
 2 AJ         0.571    -0.143     0.757     0.644     0.197     0.233     4.000     3.000
 3 BA         0.353     0.294     0.676     0.297     0.120     0.120     6.000    11.000
 4 BS         0.214     0.571     0.676    -0.007    -0.002    -0.002     3.000    11.000
 5 CC         0.231     0.538     0.486    -0.478    -0.148    -0.176     3.000    10.000
 6 CD         0.300     0.400     0.527    -0.289    -0.073    -0.095     3.000     7.000
 7 CE         0.200     0.600     0.703    -0.057    -0.015    -0.016     2.000     8.000
 8 CH         0.556    -0.111     0.608     0.351     0.145     0.157    10.000     8.000
 9 CJ         0.462     0.077     0.568     0.105     0.033     0.040     6.000     7.000
10 CM         0.150     0.700     0.554    -0.409    -0.149    -0.149     3.000    17.000
```

Various measures of the similarity of ego's ethnic identity to that of his/her in-neighbors are given. The percentage of neighbors who have the same ethnic identity as ego (e.g. 57.1% for AJ) is the most obvious (column 1). The numbers of neighbors who have the same ethnic identity as ego (column 7) and different ethnic identity as ego (column 8) are displayed as well. The EI index (column 2) is the difference between the numbers of ties to actors outside ego's group less the number of ties to actors inside ego's group, divided by the total number of ties. Positive values, therefore, show a preponderance of "external" ties and negative values show a preponderance of "internal" or homophilous ties. A variety of other measures are provided that may be of interest, depending on the problem.

Unfortunately, as of this writing, UCINET does not have a tool for computing measures of egonet homophily for continuous attributes. For example, there is no simple tool for calculating an index of how similar ego's score on group participation is to the participation scores of those who identify ego as an acquaintance. If this sort of homophily is important, one can recover the average score of ego's alters on a continuous attribute using *Network>Ego Networks> Egonet Composition>Continuous Alter Attributes*, and calculate the difference of ego's score from the mean of ego's neighbors.

In this rather lengthy section we've looked at some ways of indexing how each actor is embedded in both the global, and their local (ego) network. How connected ego is, how central they are, whether or not their neighborhood is highly clustered and/or well connected to the global network may all be important structural aspects of an actor's location that affect their attitudes and behavior.

Of course, it is not just being connected that matters. It may matter to whom one is connected. Indexing the attributes of the others in an actor's neighborhood (composition), and measuring how similar an actor is to his/her neighbors (homophily) may also matter in predicting an actor's attitudes and behavior.

6.2.4 Independent Variables: Weights for Network Autoregression and Autocorrelation

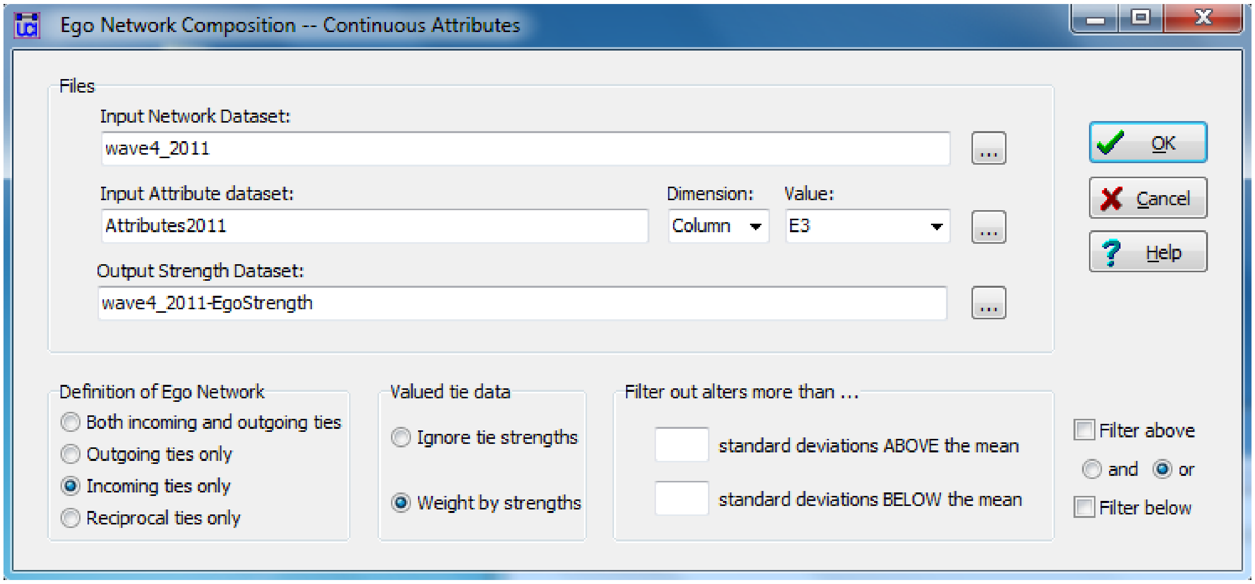
The attributes of those to whom an actor is connected may affect the actor's attitudes and behaviors. Probably the most critical attributes of ego's neighbors are the attitudes or behaviors that we want to understand. A key variable in network models of social influence then, is the values of the dependent variable for those who are connected to an actor. If we are trying to understand a student's performance on the final exam, an important predictor may be the performance of those who are most influential via the student's social network.

An actor's score on the dependent variable may be influenced by the scores of their network neighbors via two different processes: autoregression and autocorrelated error. When the scores on the dependent variable of ego's neighbors directly cause ego's score we have "autoregression." As in social influence or diffusion models, the scores of alters are treated as independent variables. Sometimes ego's score on the dependent variable may be correlated with that of ego's neighbors because of local disturbances or omitted variables. These kinds of error processes are called "autocorrelation." Autocorrelation is treated as spatially correlated error, similar to time-correlated error in time series analysis.

Whether the process involved is autoregression or autocorrelation (or both), we need to represent the scores of alters on the dependent variable as a predictor or ego's scores in network influence models. There are a variety of approaches.

First, one can simply summarize the scores of alters on the dependent variable directly. If we are only concerned with alters who are adjacent to ego, we can use *Network>Ego Networks>Egonet Composition> Continuous Alter Attributes* or *Network>Ego Networks>Egonet Composition>Categorical Alter Attributes* to create a new variable that is the mean score of alters directly tied to ego on the dependent variable, or the proportion of ego's direct-tie alters that have a particular score on the dependent attribute. Figures 6.11 and 6.12 calculate the average score of direct-tie alters on the final exam for each ego, which we will use as an independent variable to predict ego's score.

Figure 6.11. UCINET Dialog Used to Create Alters' Average Score on the Final Exam



Ego Network Composition -- Continuous Attributes

Files

Input Network Dataset: wave4_2011

Input Attribute dataset: Attributes2011 Dimension: Column Value: E3

Output Strength Dataset: wave4_2011-EgoStrength

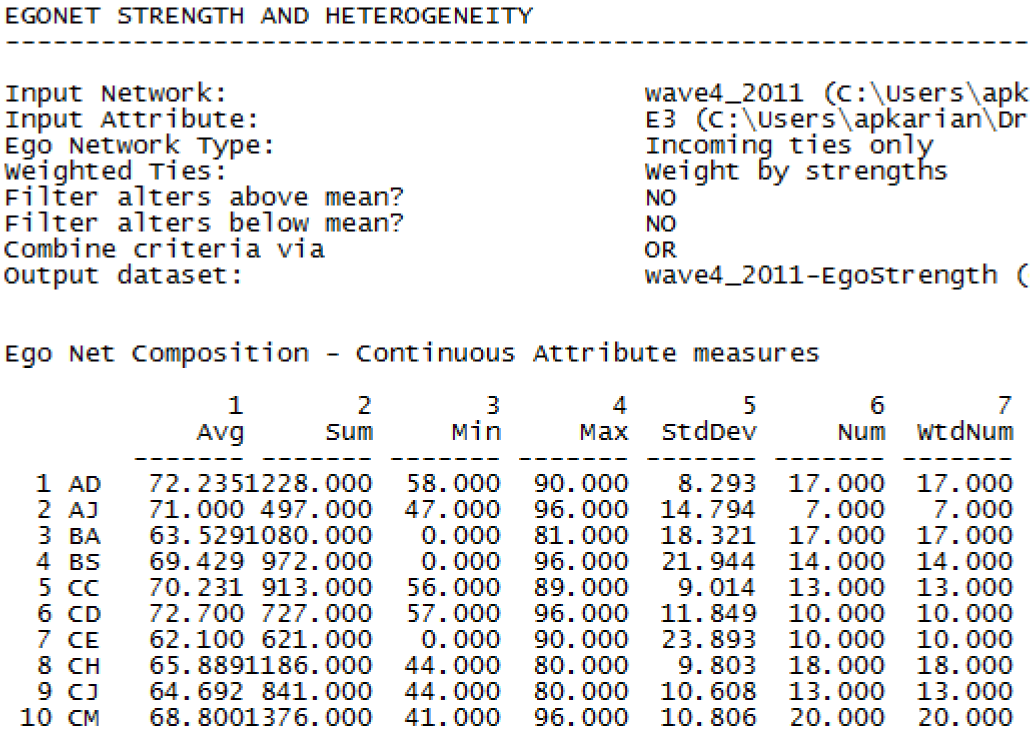
Definition of Ego Network
○ Both incoming and outgoing ties
○ Outgoing ties only
● Incoming ties only
○ Reciprocal ties only

Valued tie data
○ Ignore tie strengths
● Weight by strengths

Filter out alters more than ...
standard deviations ABOVE the mean
standard deviations BELOW the mean

□ Filter above
○ and ● or
□ Filter below

Figure 6.12. UCINET Output for Alters' Average Score on the Final Exam



```
EGONET STRENGTH AND HETEROGENEITY
----------------------------------------------------------------

Input Network:                           wave4_2011 (C:\Users\apk
Input Attribute:                         E3 (C:\Users\apkarian\Dr
Ego Network Type:                        Incoming ties only
Weighted Ties:                           Weight by strengths
Filter alters above mean?                NO
Filter alters below mean?                NO
Combine criteria via                     OR
Output dataset:                          wave4_2011-EgoStrength (


Ego Net Composition - Continuous Attribute measures

                 1       2       3       4       5       6       7
               Avg     Sum     Min     Max  StdDev     Num  WtdNum
           ------- ------- ------- ------- ------- ------- -------
  1 AD    72.2351228.000  58.000  90.000   8.293  17.000  17.000
  2 AJ    71.000 497.000  47.000  96.000  14.794   7.000   7.000
  3 BA    63.5291080.000   0.000  81.000  18.321  17.000  17.000
  4 BS    69.429 972.000   0.000  96.000  21.944  14.000  14.000
  5 CC    70.231 913.000  56.000  89.000   9.014  13.000  13.000
  6 CD    72.700 727.000  57.000  96.000  11.849  10.000  10.000
  7 CE    62.100 621.000   0.000  90.000  23.893  10.000  10.000
  8 CH    65.8891186.000  44.000  80.000   9.803  18.000  18.000
  9 CJ    64.692 841.000  44.000  80.000  10.608  13.000  13.000
 10 CM    68.8001376.000  41.000  96.000  10.806  20.000  20.000
```

We see, for example, that AJ's seven (column 6) in-neighbors had an average score of 71.0 (column 1) on the final exam, and that CJ's 13 in-neighbors did worse with an average of score of 64.7. It might be worth noting (and possibly including as a predictor variable) the variation among the neighbors. While the neighbors of BS and CC performed similarly on the exam, on the average, the performance of BS's neighbors was much more variable (see column 5).

For many autoregressive and/or autocorrelation processes, we may believe that neighbors who are more than one step from ego might have indirect effects on ego. Usually we believe that the influence of alters declines with distance from ego, and most often we think that this influence declines rapidly with distance.

For more complex ideas about the effects of neighbors at-a-distance, it is best to create a matrix of distance weights as we did in chapter 5 for calculating the Moran and Geary network autocorrelation. With a little matrix algebra and a distance-weights matrix, one can create an independent variable to use directly in modeling. Or, one can use the weights matrix in software designed for autocorrelated error and autoregression, as we do in an example later on in this chapter.

One common metric for social network "nearness" (or distance weighting) is the reciprocal of the geodesic distance between nodes. See Figure 5.11 (last chapter) to see these dyadic data nearness weights.

We now have all the pieces in place. Our dependent variable is a measure of some attribute of ego. Our independent variables include other attributes of ego, specifically, measures of the ways in which ego is embedded in the network globally and locally. We also use measures of the composition of ego's neighborhood and ego's homophily with the alters in ego's direct-tie neighborhood as independent variables. Finally, we include measures of direct-tie alter's scores on the dependent variable to measure any effects of autoregression and/or autocorrelation.

This sounds complicated. However, once the data are assembled, the analysis is an exercise in linear modeling, with the use of permutation to test hypotheses. Let's look at a variety of approaches.

6.3 Generalized Linear Models for Network Influence

As we saw above, we can create measures for each actor that describe variation in their position in the network and the influences operating on them through their network connections. For example, an actor's betweenness centrality might be used to get at the idea that actors who are more central in the global network are more likely to have more favorable outcomes because they have more dependent alters that they can draw on for resources. Or, actors who are embedded in local neighborhoods that contain many individuals who are similar to themselves and/or have high closure may perform more poorly because of the lack of diversity in the social capital that they have available.

Having created variables for each actor that measure not only their attributes, but also their structural location in the network and the influences operating on them, we can apply regular generalized linear modeling techniques to explain variation in individual attributes. This approach to understanding network influence on individual attributes has the advantage of being able to deal with actor outcomes that have a variety of distributional forms and a variety of link functions to the predictors.

Analyzing network influence this way treats each actor in the network as a case, but the cases are obviously not independent of one another. Consequently, permutation or other re-sampling methods should be used to assess the reliability of parameters. Let's look at some examples.

Suppose that we are trying to predict a student's performance on the final examination. We might treat the outcome (the individual's performance on the final exam) in a variety of ways. Since a score of 70 or more was considered to be "passing," we might dichotomize the outcome. If we took this approach, then a GLM with a binary logit or probit form might be useful. We might choose to simply rank students from best to worst exam performance. If that was our approach, then an ordered logit or probit might be useful. Or we might use all of the information available and analyze the interval-ratio total points earned on the exam, assuming a Gaussian, or perhaps log-normal or gamma distribution.

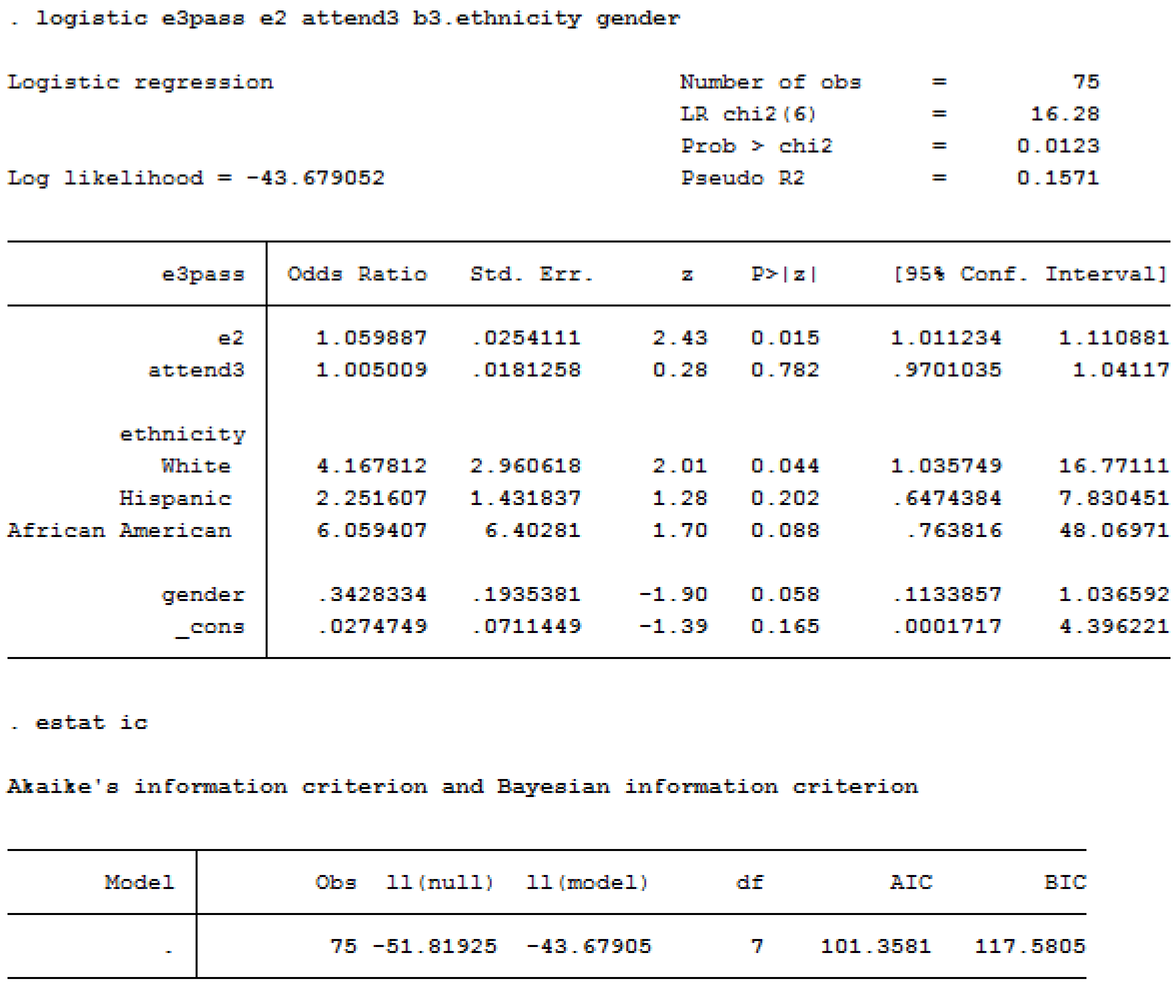
Let's look first at some results for simple binary logistic regression predictions of the odds that an individual achieved a score of 70 or higher on the final exam. In Table 6.1, several models are presented that begin with individual actor attributes, then add predictors describing global and local network position effects, and finally add effects of network influence. Figure 6.14 shows the Stata syntax and output used to generate model 2. Figure 6.15 shows the Stata syntax and output used to test the significance of the coefficients in model 2 by Monte Carlo permutation. To run these analyses, we created a standard attribute data set (actors by attributes), and used Stata's logistic regression command, followed by the post-estimation "estat" command to get information criteria measures for the models. Stata's permute command was used to generate tests of significance for the model coefficients (with 5000 replications). Asian was used as the reference category for ethnic identity because it was the modal category. Odds ratios are shown in Table 6.1.

Table 6.1. Logistic Regression Models Predicting Passing the Final Exam

Effect/Model	1	2	3	4	5
Academic Effects					
Midterm	1.06*	1.06*	1.06*	1.11**	1.12**
Attendance	1.00	1.01	1.00	0.95†	0.93*
Demographic Effects					
White		4.17*	2.96	135.79**	2206.56**
Hispanic		2.25	1.58	4.04	10.43*
African American		6.06†	4.44	668.83**	22812.40**
Woman		0.34†	0.35	0.07**	0.001**
Network Position Effects					
Degree			1.00	0.34	0.50
Betweenness			1.00	0.98	0.97†
Closure			0.95	1.08	0.97
Brokerage			1.02	1.10	1.08
Egonet Influence Effects					
Ethnic diversity				1.16**	1.28**
% Women				0.95	0.94
Gender diversity				1.10	1.19*
% passed				0.78**	0.73**
Avg. Attendance				0.71**	0.57**
Homophily Effects					
Ethnicity homophily					1.02
Gender homophily					1.11**
Pseudo R^2	0.07	0.16	0.19	0.54	0.62
AIC	102	101	106	79	74
BIC	109	118	131	116	116

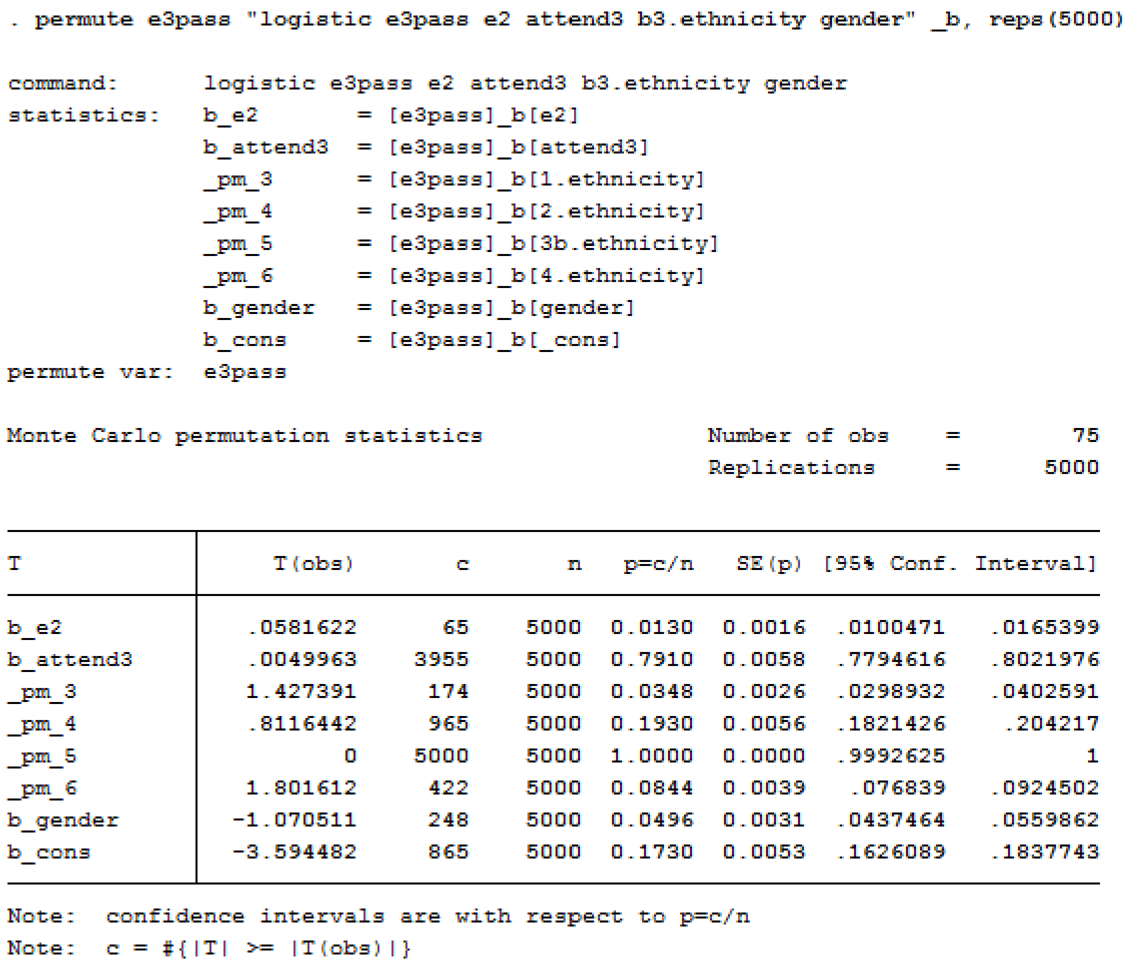
†p < 0.10; *p < 0.05; **p < 0.01, two tail, by permutation trials

Figure 6.13. Stata Syntax and Output for Logistic Regression of Passing Final Exam (Model 2)



```
. logistic e3pass e2 attend3 b3.ethnicity gender

Logistic regression                               Number of obs     =         75
                                                  LR chi2(6)        =      16.28
                                                  Prob > chi2       =     0.0123
Log likelihood = -43.679052                       Pseudo R2         =     0.1571

------------------------------------------------------------------------------------
            e3pass | Odds Ratio   Std. Err.      z    P>|z|     [95% Conf. Interval]
-------------------+----------------------------------------------------------------
                e2 |   1.059887   .0254111     2.43   0.015     1.011234    1.110881
           attend3 |   1.005009   .0181258     0.28   0.782     .9701035     1.04117
                   |
         ethnicity |
            White  |   4.167812   2.960618     2.01   0.044     1.035749    16.77111
         Hispanic  |   2.251607   1.431837     1.28   0.202     .6474384    7.830451
 African American  |   6.059407    6.40281     1.70   0.088      .763816    48.06971
                   |
            gender |   .3428334   .1935381    -1.90   0.058     .1133857    1.036592
             _cons |   .0274749   .0711449    -1.39   0.165     .0001717    4.396221
------------------------------------------------------------------------------------

. estat ic

Akaike's information criterion and Bayesian information criterion

-----------------------------------------------------------------------------
       Model |        Obs  ll(null)  ll(model)      df         AIC        BIC
-------------+---------------------------------------------------------------
           . |         75 -51.81925  -43.67905       7    101.3581   117.5805
-----------------------------------------------------------------------------
```

Figure 6.14. Stata Syntax and Output for Permutation Tests of Logistic Regression Parameters (Model 2)



```
. permute e3pass "logistic e3pass e2 attend3 b3.ethnicity gender" _b, reps(5000)

command:      logistic e3pass e2 attend3 b3.ethnicity gender
statistics:   b_e2       = [e3pass]_b[e2]
              b_attend3  = [e3pass]_b[attend3]
              _pm_3      = [e3pass]_b[1.ethnicity]
              _pm_4      = [e3pass]_b[2.ethnicity]
              _pm_5      = [e3pass]_b[3b.ethnicity]
              _pm_6      = [e3pass]_b[4.ethnicity]
              b_gender   = [e3pass]_b[gender]
              b_cons     = [e3pass]_b[_cons]
permute var:  e3pass

Monte Carlo permutation statistics                  Number of obs    =        75
                                                    Replications     =      5000

------------------------------------------------------------------------------
T            |     T(obs)       c       n   p=c/n   SE(p) [95% Conf. Interval]
-------------+----------------------------------------------------------------
b_e2         |   .0581622      65    5000  0.0130  0.0016  .0100471   .0165399
b_attend3    |   .0049963    3955    5000  0.7910  0.0058  .7794616   .8021976
_pm_3        |   1.427391     174    5000  0.0348  0.0026  .0298932   .0402591
_pm_4        |   .8116442     965    5000  0.1930  0.0056  .1821426    .204217
_pm_5        |          0    5000    5000  1.0000  0.0000  .9992625          1
_pm_6        |   1.801612     422    5000  0.0844  0.0039   .076839   .0924502
b_gender     |  -1.070511     248    5000  0.0496  0.0031  .0437464   .0559862
b_cons       |  -3.594482     865    5000  0.1730  0.0053  .1626089   .1837743
------------------------------------------------------------------------------
Note:  confidence intervals are with respect to p=c/n
Note:  c = #{|T| >= |T(obs)|}
```

In the first model, we predict passing the final exam based on whether the individual passed the mid-term, and their attendance at lecture between the mid-term and the final. The coefficients, which indicate how unit changes in the predictor multiply the odds of passing the exam, demonstrate that passing the mid-term exam has positive effects on passing the final exam. From the permutation trials, we find that effects of the size observed here occur in less than 5% of randomly permuted networks.

The second model adds the individual attributes of ethnicity and gender. Both ethnic identity and gender appear to affect the likelihood of passing the final exam for this social networks class.

The third model includes measures about the embedding of ego in the global and their local network (degree, centrality, ego-network closure, and brokerage within the ego-net). These measures were generated using the procedures shown in figures 6.1 through 6.4 above. The degree variable is the in degree of each actor, which can be found using either the “Indeg” variable from the “wave4_2011-cent” data set created in figures 6.1 and 6.2, or the “size” variable from the “wave4_2011-EgoNet” data set created in figures 6.3 and 6.4. These are identical measures. To measure centrality, we chose to use the global betweenness centrality variable (“Between”) from the “wave_2011-cent” data set. To measure ego-network closure, we reverse coded the reach efficiency (“ReachE”) variable from the “wave4_2011-EgoNet” data set because high values of ReachE indicate low levels of closure (this can be reverse coded by multiplying by negative one or subtracting from the max value). Finally, we used the brokerage variable (“Broker”) from the “wave4_2011-EgoNet” data set. None of these variables appear to be strong predictors of passing the final.

The fourth model adds social influence variables that describe the composition of each student’s ego network. Specifically, we were interested in ego’s in-neighborhood, or the direct ties that claim to know ego and therefore may influence ego’s behaviors and attitudes. The ethnic diversity variable used was from the “wave4_2011-EgoComposition-Ethnicity” data set created via the procedures shown in figures 6.5 and 6.6. We used Blau’s measure of heterogeneity (“Hetero”) multiplied by 100. The same procedure was run to generate variables describing the gender composition of the ego-networks. We multiplied the output variable called “p2,” which is the proportion of category two for the gender variable (woman), by 100 to generate a “percent woman” variable. This is the percentage of ego’s direct ties that are women. We also used Blau’s heterogeneity measure (“Hetero”) multiplied by 100 to create a gender diversity variable. In Chapter 5, we created a dichotomous variable that measured whether or not each student passed the final exam (where we’ve defined passing as achieving 70% or greater). Following the same procedure used to create the “percent woman” variable, we generated a “percent passed” variable which tells us the percent of ego’s direct ties that passed the final exam. Finally, using the same procedure outlined in figures 6.7 and 6.8, we generated a variable that examines the average attendance score (the UCINET output variable name is “Avg”) for the final third of the term (the time period most relevant to the final exam) for ego’s direct ties. Interestingly, diversity in the ethnic identities of students’ direct ties significantly improves their odds of passing, however, students who have friends with higher mid-term scores and better attendance are LESS likely to pass the final exam!

The final model tests the effects of homophily. That is, what are the effects on individuals of being in a personal network that is mostly of the same gender or same ethnicity as themselves? The creation of a measure of ethnicity homophily was demonstrated in Figures 6.9 and 6.10. This variable (“PctHomophilous”) saved in the data set “wave4_2011-EgoHomoMeas-Ethnicity” was used along with a gender homophily variable created in the same manner (both were multiplied by 100 to convert to percentages). We note a tendency (net of all the other factors) for students who have friends that are mostly the same gender as themselves to have improved chances of passing the final exam.

If we had measured the outcome as the interval-ratio variable of score on the final exam, we could use a different version of the GLM. Table 6.2 shows the results of a parallel analysis using a Gaussian distribution and identity link function (i.e. classical OLS linear regression), with permuted significance tests.

Table 6.2. OLS Linear Regression Models Predicting Final Exam Score

Effect/Model	1	2	3	4	5
Academic Effects					
Midterm	0.40**	0.32*	0.33*	0.32*	0.30*
Attendance	0.04	0.07	0.04	0.05	0.06
Demographic Effects					
White		9.13*	7.71†	12.42**	10.10†
Hispanic		3.99	2.25	5.19	4.23
African American		-2.30	-4.10	6.21	0.77
Woman		-6.24†	-5.70	-6.26†	-8.07
Network Position Effects					
Degree			0.12	2.00	2.09
Betweenness			-0.02	-0.08	-0.10
Closure			-0.15	-0.32	-0.28
Brokerage			0.10	-0.03	-0.02
Egonet Influence Effects					
Ethnic diversity				0.05	0.04
% Women				-0.12	-0.16
Gender diversity				-0.00	-0.10
% passed				-0.53**	-0.51**
Avg. Attendance				0.51	0.54
Homophily Effects					
Ethnicity homophily					-0.15
Gender homophily					0.13
R^2	0.14	0.29	0.32	0.53	0.58
AIC	595	589	593	569	565
BIC	602	605	619	605	607

†p < 0.10; *p < 0.05; **p < 0.01, two tail, by permutation trials

The analyses here are for illustration, and shouldn't be taken very seriously (see the impossibly large odds ratios in the final column of Table 6.1). We are over-fitting the data with a model that is too complex. There are some substantial collinearities among the predictors. Most important, of course, there is no well-developed theory underlying the inclusion of terms. Do note, however, that the addition of measures of how ego is structurally embedded (models 3 and beyond), and social influence (model 4 and 5) do add considerably to our ability to predict the outcome, compared to a model based on individual attributes alone. However, from the BIC, we can see that this extra explanatory power comes at a cost.

Our approach to the effects of structural embedding and social influence are rather ad-hoc. In the next section, we will examine one more theoretically grounded approach to the study of social influence and diffusion processes based explicitly in exponential random graph theory (which we will discuss at some length in upcoming chapters).

6.4 Autologistic Actor Attribute Models

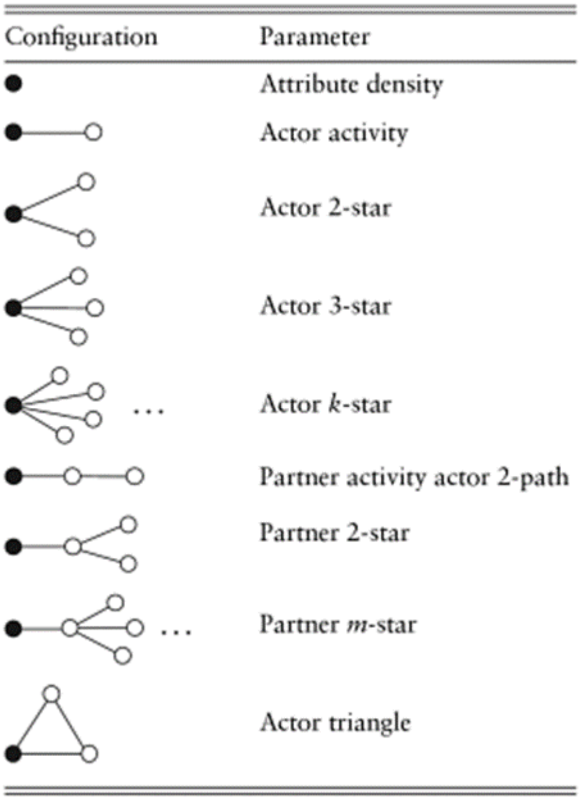
The "autologistic actor attribute model" (AAAL) is a distinctive approach to examining processes of diffusion and influence that fits in the general "exponential random graphs" (ERG) framework that we'll examine more closely in the next few chapters. An excellent recent presentation of the general approach to the AAAL is contained in Lusher, et. al (2013). A formal presentation of the model itself is given in Draganova and Robins (2013), and an illustration of its use to study network effects on employment status is provided in Draganova and Pattison (2013). The first of these works provides a very nice literature review that situates the AAAL within the ERG literature, and discusses precursors and competing approaches. Software for estimating AAAL (called iPNet) is available from http://www.melnet.org.au/pnet/, which is the home of the research group that has been a leading innovator in this field for many years.

At this writing, the AAAL model software supports the analysis of binary actor attributes (i.e. there are no versions for ordinal, multinomial, or interval-ratio attributes), and embedding in a symmetric binary network. The dependent variable is the presence or absence of the attribute for each ego, and ego's attributes may be used to predict the log-odds of the presence of the attribute, just as we might if we were treating each ego as an independent observation.

The AAAL model, though, allows us to estimate a number of different kinds of social influence effects based on characteristics of ego's one-step ego network. There are three general classes of social influence effects of this type: "network position" effects; "network-attribute" effects; and "covariate" effects.

"Network position effects" characterize some important aspects of the structure of each ego's neighborhood, independent of attributes of the neighbors. These network position effects represent hypotheses about how the embedding of an ego in his/her local network may affect the probability that they have the attribute. Figure 6.15 (adapted from Daraganova and Robins 2013, table 9.1) graphically illustrates the types of network position effects that can be estimated with the AAAL. Note that the circles representing the alters directly or indirectly tied to ego (the shaded circle) are left unshaded which indicates that the presence or absence of the alters' attribute is not important for these types of effects.

Figure 6.15. AAAL Network Position Effects



Configuration	Parameter
●	Attribute density
	Actor activity
	Actor 2-star
	Actor 3-star
. . .	Actor *k*-star
	Partner activity actor 2-path
	Partner 2-star
. . .	Partner *m*-star
	Actor triangle

The first five effects form a hierarchy that models the majority of the variation in the degree-distribution of the graph. That is, these effects examine the hypothesis that egos that have fewer (or more) alters are less (or more) likely to exhibit the trait. Suppose that we were interested in predicting whether our students passed the final exam. We might think that students who had more acquaintances had more information, study partners, and social support, and would be more likely to pass the exam (regardless of their own attributes, or the attributes of their partners). The "actor activity" effect codes whether each ego has one (or more) alters. The "actor 2-star" effect codes whether an ego has two (or more) alters (controlling for whether they have any partners). The "actor 3-star" and "k-star" further differentiate actors who have still more alters from those who have none or fewer. Most degree distributions of actors are steeply exponential, so knowing whether actors have any alters or not, or knowing if they have more than one or two partners usually accounts for most of the variation in degree distributions.

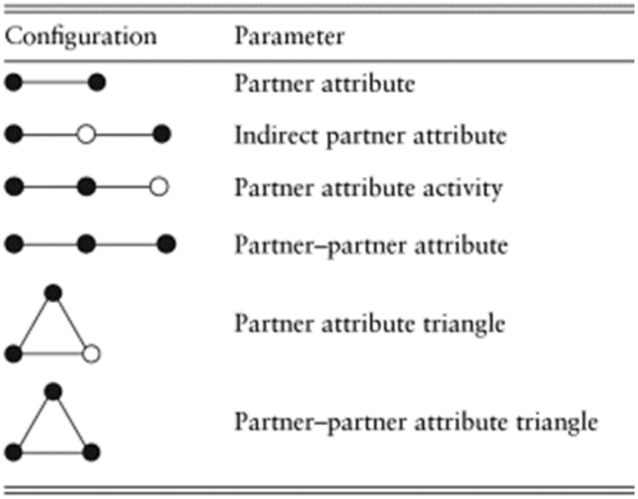
The next three network position effects in figure 6.15 capture the idea that some actors are connected to others who are themselves well connected, while other egos may have equal numbers of partners, but those partners are less well connected. This is the notion of eigenvector or power/influence centrality. We might suppose, all else equal, that actors who are connected to well-connected others have easy access to large quantities of support, information, and influence. In our example, we might suppose that students who are well embedded in the "in-crowd" are in an advantaged position in preparing for the final exam.

The last network position effect in figure 6.15 is the "actor triangle." Given that an ego has at least two alters, this effect asks: are these alters connected to one another? This gets at the idea of "clustering," or "closure," or lack of "structural holes" in ego networks. Actors who are tied to other actors who are also tied together (net of other structural effects) may have less efficient networks. Speculatively, we might suppose that students with neighborhoods characterized by "cliques" or closed structures may do more poorly on the examination because they have access to less unique information and perspective per social tie.

The network position effects of the AAAL are intended to capture the most important structural aspects of the ego-neighborhoods of each node (degree distribution, centrality, closure). They lead us to think about why some actors might be more likely to display an attribute based solely on how they are connected, regardless of the attributes of those to whom they are connected.

The next class of effects in the AAAL model, "network-attribute" effects, model specific forms of local autoregression. Figure 6.16 (again, adapted from Daraganova and Robins 2013) illustrates the effects available in iPNet. Notice now that certain alters directly or indirectly tied to ego are shaded, indicating that the presence or absence of those alters' attribute *is* important for these types of effects.

Figure 6.16. AAAL Network-attribute Effects



Configuration	Parameter
●—●	Partner attribute
●—○—●	Indirect partner attribute
●—●—○	Partner attribute activity
●—●—●	Partner–partner attribute
	Partner attribute triangle
	Partner–partner attribute triangle

The most obvious, and often most important autoregressive effect is the "partner attribute" effect. It hypothesizes that the presence or absence of the dependent attribute for alter affects the probability that ego also has the attribute. The AAAL model also includes additional possible effects of the prevalence and location of the attribute in ego's neighborhood on the likelihood that ego displays the attribute.

The "indirect partner attribute" hypothesizes that if ego's direct-tie alter has an alter that has the attribute (regardless of whether ego's direct-tie alter does), this indirect influence may affect ego's outcome. The "partner attribute activity" effect hypothesizes that ego's direct-tie alter may be more likely to lead ego to adopt the attribute if the direct-tie alter has alters (we are more influenced by others who are popular). This partner attribute activity effect may be even stronger to the extent that the direct-tie alter's alters also display the attribute ("partner-partner attribute"). Again, as with network position effects, note that the network attribute effects are hierarchical – one must have effects from partners in order to also possibly have effects of partner's partners. These effects suggest, for example, that the performance of the student's friend's friends on the exam may influence our focal student's chances.

The last two effects in figure 6.16 combine the closure of ego's neighborhood, and the prevalence of the attribute in that neighborhood. The "partner-attribute triangle" suggests that being in a closed neighborhood where anyone else has adopted the attribute may affect the likelihood of adopting. The "partner-partner attribute triangle" suggests that if ego is embedded in a clique where everyone else has the attribute, they are also likely to display it. If a student is in a clique where anyone passes the exam, or where everyone else passes the exam, they may themselves be more likely to pass the exam.

Finally, the AAAL model provides ways of modeling the effects of other attributes on the likelihood that ego displays a trait ("covariate effects"). Obviously, we might suppose that ego's own attributes affect the likelihood that they display an attribute. For example, students who have done well on previous exams may be more likely to pass the final.

It also might be true that students who are tied to others with certain attributes may have different outcomes, regardless of their own attributes. In AAAL, these are called "partner-covariate" effects. For example, a student who is connected to others who did well on previous exams might be more likely to do well on the final.

Whether ego and alter are homophilous on other attributes may also by hypothesized to affect ego's outcomes. In AAAL, such homophily effects are termed "same-partner-covariate" effects. If a man has a bias toward other men in his acquaintanceship network, he may be more likely to succeed (e.g. do well on a final examination) because of the sense of security and social support and trust that may be more common in homophilous ego-networks.

Taken together, the AAAL model provides a powerful tool-kit for examining social influence processes. It is a particularly interesting approach because it identifies and suggests hypotheses about effects on outcomes of the purely structural aspects of how actors are embedded in networks. It also includes a rich approach to understanding autoregressive effects, and the effects of other attributes of both ego and alter.

The iPNet software is relatively easy to use. It is available free for download, and is easy to install.

iPNET example using student data coming soon

6.5 Autoregressive and Error Correlation Regression Models

So far, we've considered some ways to estimate the effects that the attributes of others have on an actor's own attributes. In this section, we'll look at ways in which we can control for special kinds of network influence effects that may act as nuisances in our models predicting attributes: network autoregressive and spatial error processes.

Network autoregression exists when ego's score on an outcome attribute is determined (or at least predicted) by alter's score on the same attribute. For example, we might hypothesize that a student's score on the final exam is caused, or predicted, by his/her neighbor's (in the ego-net) scores on the final exam. This kind of effect can be included in GLM (see section 6.3 above) and AAAL models. If ego's neighbors do, in fact, exert social influence on ego, then it seems reasonable that ego may have outcomes that are more similar to his/her alters than to random others in the network. In GLM and AAAL models, we are trying to directly model the attributes of the alters that may affect ego's outcome. If evidence of such effects exist, then there is autoregression in the network.

Network autoregressive models seek to control for these types of network influences, with the main goal of correctly estimating the effects of ego's own attributes on ego's outcomes. Unlike the GLM and AAAL approaches, network autoregressive models don't try to directly examine the network influence processes, but simply remove these potentially confounding effects. Network autoregressive models provide a convenient way of controlling for network autoregressive influences of actors at greater network distances. If our interest is primarily in how ego's attributes are associated with ego's outcomes, we may well wish to simply control for, rather than directly model the effects of the alters' outcome scores on ego's outcome scores.

Network autocorrelated error exists when there are unmeasured local variables (errors, disturbances) that may affect the scores of both ego and ego's alters on an outcome variable. Suppose that one student, while trying to prepare for the final examination, had a family crisis, and called on his/her friends in the class for support. We might suppose that ego's score on the final exam would be lower than we would have expected on the basis of measured variables because of the "local disturbance" of having his/her study interrupted. But, we might expect that ego's friend's scores will also be negatively affected as they are deflected from preparing for the exam by supporting ego. In this case, both ego and his/her connected alters end up studying less for the final exam than we would have expected on the basis of their own attributes, resulting in a correlation of the error terms for ego, and the error terms for the alters.

Autoregressive and autocorrelated error models originate from geo-spatial analyses. In geo-spatial analysis, scores on some outcome (say, crime rates) are likely to be similar in places that are geographically close to one another. There are several processes that produce these spatial correlations. Because of exogenous processes (say the operation of the social class system of society), the attributes of actors who are spatially close are likely to be similar, producing similar outcomes (without any influence processes at all). The scores of spatially adjacent actors may also be similar because they are influencing one another. Criminals who practice their craft in one neighborhood are likely to also seek targets in adjacent spaces. But, there may be additional local disturbances that produce more similarity in crime rates in adjacent areas than we would otherwise expect. Perhaps we failed to measure and control for the level of policing, which is likely to be similar in adjacent neighborhoods. Having ignored this important variable will produce similar errors of prediction in adjacent neighborhoods.

To solve this problem, we simply apply the social network distance between the observations, and use the same approaches as geographical analysis, but with network distance, rather than spatial distance.

The network autoregressive process is also known as the "spatial lag model". That is, the score on the outcome attribute of ego depends on the outcome attribute of ego's alters, in addition to ego's own attributes.

The network autocorrelated error model is also known as a model with "autoregressive disturbances". In this model, the residuals or prediction errors of ego are correlated with the residuals or prediction errors of the alters.

It is possible, of course, to suppose that both lag and correlated error processes are operating.

To estimate models that correct for network auto regressive and/or autoregressive disturbances, the *spreg* package (Drukker, et al.) for *Stata* is rather easy to use. Pisani (2012) describes *spreg* as follows:

> spreg estimates the parameters of a cross-sectional spatial-autoregressive model with spatial-autoregressive disturbances which is known as a SARAR model. A SARAR model includes a weighted average of the dependent variable, known as a spatial lag, as a right-hand-side variable and it allows the disturbance term to depend on a weighted average of the disturbances corresponding to other units. The weights may differ for each observation and are frequently inversely related to the distance from the current observation. These weights must be stored in a spatial-weighting matrix created by spmat. spreg estimates the parameters by either maximum likelihood (ML) or by generalized spatial two-stage least squares (GS2SLS).

Detailed documentation of the Stata packages needed for these types of models (also known as Cliff-Ord models) is available from Drukker, Peng, Prucha, and Raciborski (2013) and Drukker, Prucha, and Raciborski (2013). The packages are used to create spatial weighting matrices (*spmat*) and perform Cliff-Ord regressions (*spreg*).

Let's again consider the problem of predicting our student's scores on the final examination. We've already created the necessary input to do the analysis. In the section on GLM models (above), we created an attribute data set for our 75 students that includes individual attributes, measures of the location of each student in the global network (degree and betweenness centrality) and local neighborhood (closure and brokerage), and social influence (attributes of ego's neighbors). Previously (in Chapter 5) we created a network distance weights matrix in order to test for autocorrelation. This is a student-by-student matrix of the reciprocal of geodesic distances. Of course, different definitions of distance could be used (for some ideas, see Chapter 5), and different distance matrices could be used for the autoregressive and error terms (we will use the same for both).

After locating and installing the *sppack* package (STATA: *findit sppack*), we locate our attribute file and distances file in a working directory (note the path and name). Estimation is a two-step process. First, we use the distance weights file we created last chapter (wave4_2011-geo.dta), and convert it to a form that can be used by the spatial regression package. We then load the attribute file (created for section 6.3 above), and perform regressions. Figure 6.17 gives the edited syntax of the STATA .do (i.e. batch) file that we used for this example (syntax for models 5 and 6 have been removed for brevity). Note, data files should be saved as .dta files prior to running. Also, make sure to specify the appropriate directory needed to find the data files.

Figure 6.17. STATA Syntax Used to Create a Distance Matrix and Perform Several Cliff-Ord Regressions

```
* Call in inverse geodesic distance weights
use "C:\SARAR\ wave4_2011-geo.dta", clear

* Create a spatial-weighting matrix object
spmat dta distwt var*

* Save the spatial-weighting matrix object so that it can be called in when using attribute data
spmat save distwt using distwt.spmat, replace

* Open the attribute data
use "C:\SARAR\attributes", clear

* spreg wants the id to be numeric, so gen a new id that goes from 1-75
gen idnum = _n

*****
* You would use the following line of syntax to call in the spmat objects
* that were previously created with spmat.
* spmat use distwt using distwt.spmat
*****

* Run the regressions
* dlmat uses spatial weights for network autoregression
* elmat uses spatial weights for autocorrelated errors

* first model includes only the autoregression based on inverse geodesic distance
spreg ml e3, id(idnum) dlmat(distwt)
* get AIC and BIC
estat ic

* second model includes only the autocorrelated errors
spreg ml e3,  id(idnum)  elmat(distwt)
estat ic

* third model includes both autoregression and autocorrelated errors
spreg ml e3,  id(idnum)  dlmat(distwt)  elmat(distwt)
estat ic

* fourth model includes only covariates with no autoreg or autocorr
spreg ml e3 e2 woman part btwcnt egosize pctwoman pctwhite, id(idnum)
estat ic

* seventh model includes covariates, autoregession, and autcorelated error
spreg ml e3 e2 attend3 white hisp afam gender indeg between inv_reacheff broker heteroeth
pctwoman heterogen attendavg homoeth homogen, id(idnum) dlmat(distwt) elmat(distwt)
estat ic
```

Table 6.3 summarizes the results of seven different models predicting student's scores on the final examination. The first models (1-3) include only the autoregressive and error correlation terms (separately and then together). The fourth model includes only covariates that describe attributes of ego, ego's position in the network, and ego's local neighborhood. The variables analyzed in section 6.3 were used again here. The spreg command does not support Stata's "factor" variables so dummies were generated for ethnicity. We dropped the "percent passing the final" variable because it is, in effect, an autoregressive term. The final three models include the covariates and the autoregressive and autocorrelated error terms. The models are for instructional use only, and though they use real classroom data, they shouldn't be taken seriously as testing a well specified theory of student performance.

Table 6.3. Cliff-Ord Regression Models Predicting Final Exam Score

Effect/Model	1	2	3	4	5	6	7
Intercept	59.4**	67.3**	60.6**	34.5†	35.3*	39.0**	39.6**
Academic Effects							
Midterm				0.31*	0.32**	0.24**	0.24**
Attendance				0.08	0.09	0.09	0.09
Demographic Effects							
White				4.94	3.95	7.93*	7.38*
Hispanic				2.16	2.06	1.61	1.57
African American				-9.92	-9.39	-11.17*	-10.72*
Woman				-6.48	-7.06†	-5.69†	-6.03†
Network Position Effects							
Degree				0.63	2.09	1.75	2.29
Betweenness				-0.04	-0.01	-0.08	-0.06
Closure				0.08	-0.04	0.09	0.05
Brokerage				0.04	-0.07	-0.05	-0.10
Egonet Influence Effects							
Ethnic diversity				-0.18	-0.15	-0.23	-0.22
% Women				-0.03	-0.05	-0.00	-0.01
Gender diversity				-0.33	-0.32†	-0.31†	-0.31†
Avg. Attendance				0.38	0.47	0.35	0.38
Homophily Effects							
Ethnicity homophily				-0.20†	-0.19*	-0.24*	-0.24**
Gender homophily				0.12	0.14	0.11	0.12
SARAR Effects							
Autoregression (Lambda)	0.00		0.00		-0.01		-0.00
Autocorrelated Error (Rho)		-0.11*	-0.11*			-0.20**	-0.19**
Log-likelihood	-300	-297	-297	-281	-280	-273	-273
AIC	606	600	602	596	599	585	587
BIC	613	607	611	635	643	629	633

†p < 0.10; *p < 0.05; **p < 0.01, two tail

The significance of the coefficients taken with the goodness of fit statistics (deviance, AIC, BIC) suggest that including a network autocorrelated error term (Rho) is a useful addition to the covariates. However, accounting for autoregression (Lambda) does not appear to improve the model fit. Given the fit statistics, it appears that Model 6 might be the "best" model of the bunch.

We usually would anticipate that network autoregressive effects and network autocorrelated error effects would be positive quantities. That is, we would probably expect that being surrounded by others who do well on the exam would be associated with doing well oneself. This does not appear to be the case given the models that include Lambda. Similarly, we might expect that local disturbances (disturbances affecting an ego network, but not the whole network) would produce more similar outcomes among the members of the ego-network. However, the results above tell us just the opposite. These findings are consistent with the models in Tables 6.1 and 6.2 which find that being tied to others that pass the exam in one's egonet causes students to do worse on the exam!

6.6 Summary

In this chapter we introduced models and techniques for capturing the ways in which network processes (e.g. social learning, social influence, diffusion) can influence nodal attributes. Using the student data introduced in Chapter 2, we provided examples demonstrating how to generate measures of network position, local influence (via egonets), and homophily in UCINET. We also discussed ways of generating spatial lags for response variables and spatial weights that can be used to model processes of autoregression and autocorrelation. We also briefly introduced Autologistic Actor Attribute Models, and finally, demonstrated how to control for autoregression effects and autocorrelated errors in Stata.

6.7 References

Burt, Ronald. 1992. *Structural Holes: The Social Structure of Competition*. Cambridge, MA: Harvard University Press.

Draganova, Galina and Pip Pattison. 2013. "Autologistic Actor Attribute Model Analysis of Unemployment: Dual Importance of Who You Know and Where You Live." Pp. 237-247 in Dean Lusher, Johan Koskinen, and Garry Robins (Eds.), *Exponential Random Graph Models for Social Networks: Theory, Methods, and Applications*. Cambridge: Cambridge University Press.

Draganova, Galina and Garry Robins. 2013. "Autologistic Actor Attribute Models." Pp. 102-114 in Dean Lusher, Johan Koskinen, and Garry Robins (Eds.), *Exponential Random Graph Models for Social Networks: Theory, Methods, and Applications*. Cambridge: Cambridge University Press.

Drukker, David M., Hua Peng, Ingmar R. Prucha, and Rafal Raciborski. 2013. "Creating and managing spatial-weighting matrices with the spmat command." *The Stata Journal,* 13(2): 242–286.

Drukker, David M., Ingmar R. Prucha, and Rafal Raciborski. 2013. "Maximum likelihood and generalized spatial two-stage least-squares estimators for a spatial-autoregressive model with spatial-autoregressive disturbances." *The Stata Journal,* 13(2): 221–241.

Gould, Roger V. and Roberto M. Fernandez. 1989. "Structures of mediation: A formal approach to brokerage in transaction networks." *Sociological Methodology,* 19: 89-126.

Chapter 7. Models of Network Selection: Basics

7.1 Network as Dependent Variable

The distinctive aspect of social network analysis is its focus on the relations between actors, rather than the attributes of actors, as the basic building-block of society. What is probably most distinctive about the application of statistical modeling in SNA is its tool kit for thinking about patterns of relations (or social structures, or networks) as a dependent variable. In this chapter, we'll take a look at how SNA approaches explaining and predicting networks.

The formal mathematical definition of a network is very simply a set of nodes and the set of edges (if directed) or arcs (if non-directed) among them. The models we are considering have a fixed set of nodes. So, what distinguishes one network from another is the presence or absence of arcs/edges, or their strengths. To describe a network (or the difference between two), we would make lists. These lists would have the identities of all possible pairs of nodes, along with a description of the relation between the nodes. That is, networks can be described, formally, as lists of the "attributes" of all possible dyads.

Less formally, statistical models for explaining or predicting networks operate by using independent variables to predict the state of the relation in each possible dyad among a set of actors. If we had a group of three people (A, B, C), and the symmetric tie of "is a friend of," we would seek to build a model to explain or predict whether A and B were friends, whether A and C were friends, and whether B and C were friends. In this simple case, there are four possible "kinds" of networks or social structures (no ties, one tie, two ties, three ties). The unit of analysis when we are taking the network as our dependent variable is the dyad.

Building theories about why social structures or networks vary then, is building theories about how ties are created and ties are dissolved with different probabilities that are dependent on explanatory variables. That is, SNA theories of social structure are "generative" and talk about processes that create or break social ties. The corresponding statistical models used to describe networks or test hypotheses about them involve using independent variables to predict, for each dyad, the state of the relation between them.

The unit of observation is the dyadic tie. The "N" in models predicting networks is the number of possible dyads or unique pairs of actors. For directed networks (i.e. where AB does not necessarily imply BA), the number of dyads is *K*K-1*, where *K* is the number of actors. For non-directed networks, it is half this number (because AB implies BA).

The dependent variable (the tie between A and B) can be measured at any level. To date, most statistical models for networks use simple binary measures (the tie AB is either present = 1, or absent = 0). In principle, however, ordinal, counts, multinomial, or interval measures of ties could be used.

The independent variables are used to explain variation across dyads in the presence/absence (or form, or strength) of ties. Following Lusher and Robbins (2013, p. 24), it is useful to think of three broad classes of independent variables in models predicting networks: variables that reflect network self-organization processes, variables that measure effects of actor attributes, and variables that measure the effects of exogenous context.

Network self-organization processes suggest that the likelihood of a particular new tie forming (or an existing one disappearing) depends upon the current state of the network and the location of the tie in that network. This might sound a bit abstract, and some examples of self-organizing processes will help.

Consider three actors (A, B, C), with no ties. If a tie is added, which one will it be? It could be AB, BA, AC, CA, BC, or CB (if the ties are directed). Knowing nothing else about A, B, and C, we would probably guess that any one of the six possibilities is equally likely.

Suppose that the tie AB did form (here, A is the source node, the one directing the tie, and B is the destination node). What happens next? There are five possibilities (BA, AC, CA, BC, CB). One theory (randomness) would predict that each of these five is equally likely. But, we might have some alternative, substantive theories.

The theory of "activity" suggests that actors may differ in their capacity/motivation to direct ties at others. If this theory is true, then the tie AC might be more likely than any of the others (because an actor who has already formed one tie is more likely to form another; i.e. A's behavior can be categorized as "active" or "outgoing"). Or, we might suppose that "popularity" is operating. If actors vary in popularity, then actor B, who was selected on the first round, would be more likely to be selected by actor C forming CB on the second round. Theories of social processes suggest that some actors are more capable/likely to make ties than others (they are more active or sociable). Theories also sometimes suggest that "the rich get richer" or that actors who have ties already are more likely to get additional ones.

In many social situations, there are norms of reciprocity operating. If we theorize that norms of reciprocity are dominant in the three node network, we might then hypothesize that in the network with a single tie, AB, the next tie to form would most likely to be BA.

Alternatively, a theory of brokering might predict that in the three node network with a single tie AB, the next tie to emerge would most likely be BC, forming a line in which B is a broker between A and C.

What if two ties exist, what is likely to happen next? Popularity, activity, and reciprocity might all continue to operate. But, new kinds of structures can now emerge. If we have AB and BC, we might see CA form – creating a "closed circuit." If instead, the directed tie AC emerged, forming a triad, we'd find a more complex hierarchy.

There are a number of different ways of thinking about these sorts of self-organizing processes. Most theorists suggest that activity/popularity (or the shape of the in and out-degree distributions), reciprocity, closure, and brokerage are often important processes. Some analysts think that most of the important variation in network structures can be described by looking at configurations of three actors (see our discussion of the triad census, below). Other analysts believe that some additional more complex structures (e.g. cliques of 4 or more actors, multiple 2-paths, multiple triangles, etc.) are formed by additional social processes.

In any case, the basic notion of self-organizing network processes is that the existing structure of a network affects what happens next. In other words, the likelihood of a tie forming or dissolving depends on how it is embedded in the network structure itself. The processes being described are general ones that operate independently of the attributes of the actors.

Actor attribute processes of generating networks are more familiar. We might suppose that men are less likely than women to form ties due to gender socialization and expectations. We might suppose that actors who are supervisors are more likely to form out-ties, and that actors who are workers are more likely to form in-ties. We might also think that ties between supervisors or between workers are more likely to be reciprocated, while ties between workers and supervisors are less likely to be reciprocated. That is, we might suppose that the likelihood of a given tie forming depends on the attributes of the potential sender alone, the potential receiver alone, or the attributes of both (homophily, anti-homophily, hierarchy, etc.).

The effects of actors' attributes bias the basic processes of tie formation. That is, they "interact" with the self-organizing tendencies of networks. To say that women are likely to have more ties than men is to suggest that gender affects or biases processes that create density and degree distributions. To say that two nodes at the same level in a hierarchy (workers or supervisors) are more likely to form a reciprocal tie suggests that homophily of rank increases the probability of a tie being reciprocated. Triads and larger structures may be more likely to form among actors who share the same trait (ethnicity, gender, age, etc.).

Exogenous context, or other dyadic relations among actors may also affect the likelihood of forming ties, reciprocating ties, forming closed structures, and the like.

In our classroom example (see Chapter 2, or the "Classroom Data Codebook" found here, for details), we assigned students to work-groups – i.e. we externally imposed a structure of affiliation. We might suppose that being affiliated with the same workgroup would set processes in motion that would create a greater density of ties (and perhaps reciprocity and closure) among group members. In an organizational setting, we might suppose that the formal "chain-of-command" network would bias the network of informal "friendship" or "advice seeking" ties. Two actors who are closer together in geographical or temporal space might be more likely to form ties than actors who are further apart.

In all of these examples, one dyadic relation is shaping or "training" the pattern of ties in another. The ties of affiliation of students with workgroups would normally be represented as a student-by-student matrix, with a "1" indicating that the dyad were in the same work group, and a "0" if not. Similarly, the chain-of-command in an organization can be represented as a network of who gives orders to whom. The geographical or temporal distances among actors are represented as a dyadic matrix of distance or closeness.

To quickly summarize:

Analyzing networks (or relations or structures) as dependent variables is done by treating the state of the relation of each dyad as the dependent variable. "N" is the number of pairs of actors, or dyads.

Conventional descriptive/predictive modeling approaches are then used to account for variation across dyads in the nature of their relation. By far the most common approach is logistic regression predicting the presence or absence of each dyadic relation.

While any set of independent variables can be introduced to explain or predict dyadic relations, SNA approaches suggest that self-generating structural processes, attributes of the actors, and the effects of structural contexts (other networks) are important classes of explanatory variables.

The most important and widely used general approach to predicting or explaining networks as outcomes is the "Exponential Random Graph" (ERG) or P-star tradition. There are many local variations, adaptations, extensions, innovations, and advanced applications tuned to particular research areas in this general tradition. In this chapter, we'll first look at a simple tool in UCINET for building a basic ERG model. Then, in the next chapter, we'll briefly look at the basics of PNet, which allows a much greater range of models (but is more difficult to use).

At present, ERG modeling approaches for non-binary networks have not been developed to any useful level. So, analyzing networks that measure relations as counts, probabilities, ranks, types, or strengths are not readily available. To a limited degree, it is possible to use multi-level models in the generalized linear modeling tradition for outcomes of these types. We'll look at some ideas along these lines in the next chapter as well.

7.2 The Triad Census

The generative theories of social networks are largely "bottom-up" theories. Networks are seen as emerging from the agency of actors operating in local neighborhoods, which make and break ties. In looking at larger social structures then, it is often interesting to examine the variety of local structures that comprise them. Suppose that the social network among the students in one class has many more ties than another, but that ties in the second class are much more likely to be reciprocated if they exist. These two classes have quite different potentials for how they may behave at the macro level. At the micro level, most students in the first classroom are likely to have at least some ties to other students, but students are likely to have "open" ego-networks. In the second classroom, students have fewer alters per social tie, and form tighter and more closed local social worlds.

The exponential random graph approach to modeling networks predicts the presence or absence of a tie (or the tie strength) between each pair of actors as a function of a small number of basic network parameters, and how actor (and dyadic) attributes affect these parameters. For example, an ERG model might propose that women have more ties than men, but that the ties of men are more likely to be reciprocated.

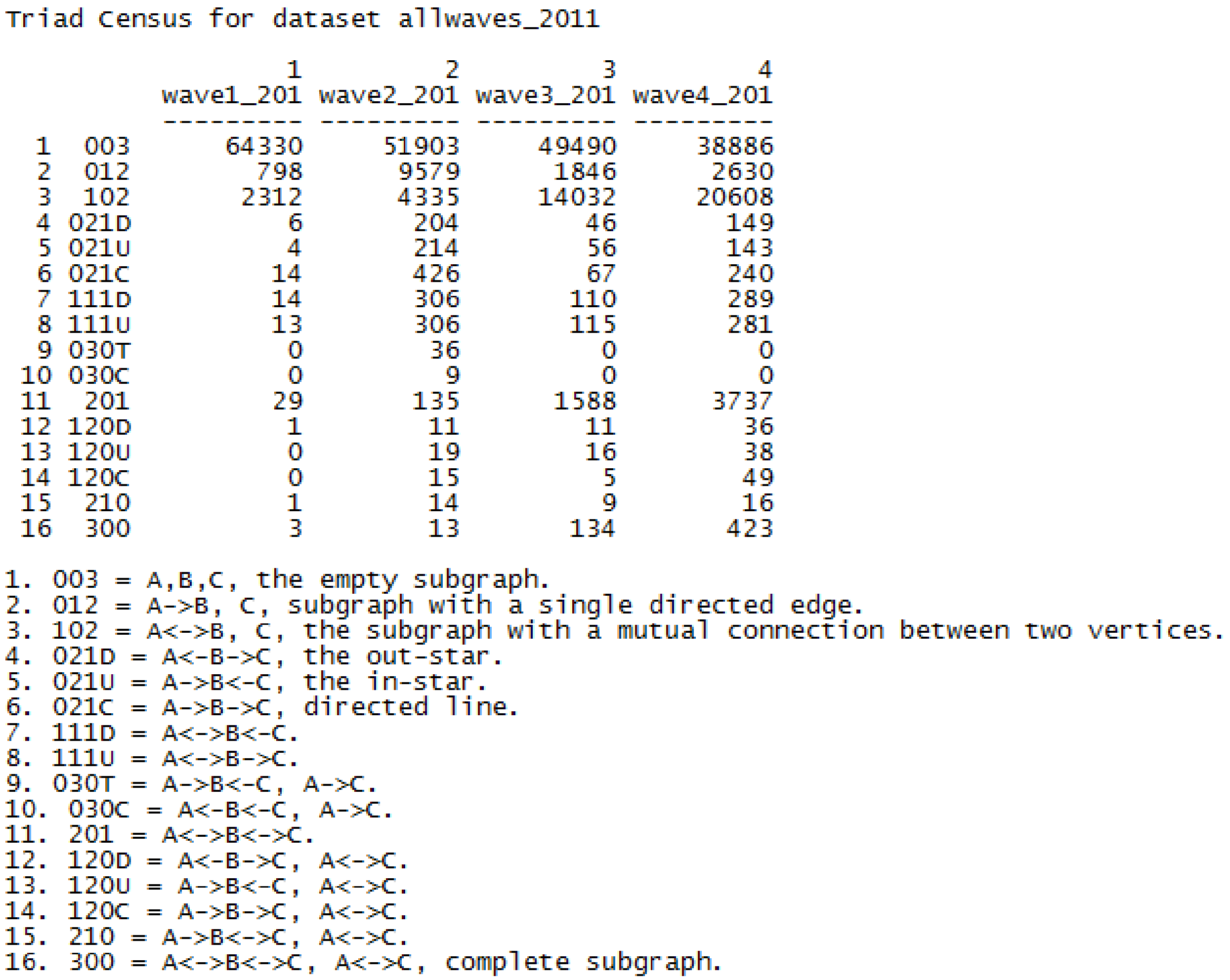
Before doing ERG modeling, it is a good idea to take a look at the prevalence of various local structures in the network. One tool for this is the "triad census," which counts up the numbers of triads in a graph that have all logically possible structures.

For symmetric (bonded, non-directed) graphs, the story of the triad census is pretty simple. Focusing on any given triad, it can have one of four possible structures: no ties, one tie, two ties, or all three ties. Obviously, the more triads there are that have multiple ties, the greater the overall density. But the ratio of the number of triads that have two ties to those that have three ties tells us something about the likelihood of transitivity for a given density, for example.

For asymmetric (directed) graphs, the story is more complicated. The number of ties among a given three actor set can vary from 0 to 6. But, more than that, there is more than one way for a directed triad to have, for example, two ties. In fact, there are 16 possible configurations of the ties among three actors. The triad census for directed data, then, reports on the prevalence of 16 different configurations.

In figure 7.1, we see the results of running UCINET's *Network>Triad Census* on all four of the waves of the student acquaintanceship data in its asymmetric, or directed form using the file "allwaves_2011" (this file was created in Chapter 2).

Figure 7.1. UCINET's Triad Census for Student Acquaintanceship, Asymmetric



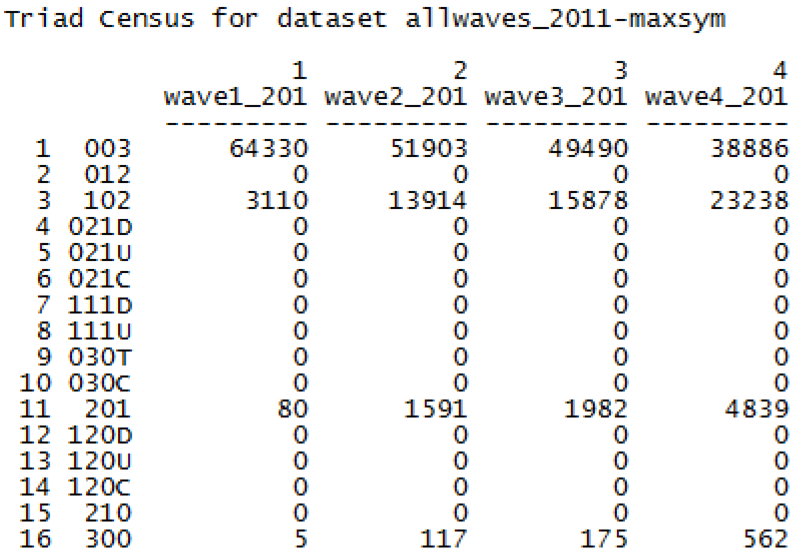
```
Triad Census for dataset allwaves_2011

                   1          2          3          4
           wave1_201  wave2_201  wave3_201  wave4_201
           ---------  ---------  ---------  ---------
  1  003       64330      51903      49490      38886
  2  012         798       9579       1846       2630
  3  102        2312       4335      14032      20608
  4 021D           6        204         46        149
  5 021U           4        214         56        143
  6 021C          14        426         67        240
  7 111D          14        306        110        289
  8 111U          13        306        115        281
  9 030T           0         36          0          0
 10 030C           0          9          0          0
 11  201          29        135       1588       3737
 12 120D           1         11         11         36
 13 120U           0         19         16         38
 14 120C           0         15          5         49
 15  210           1         14          9         16
 16  300           3         13        134        423

1. 003 = A,B,C, the empty subgraph.
2. 012 = A->B, C, subgraph with a single directed edge.
3. 102 = A<->B, C, the subgraph with a mutual connection between two vertices.
4. 021D = A<-B->C, the out-star.
5. 021U = A->B<-C, the in-star.
6. 021C = A->B->C, directed line.
7. 111D = A<->B<-C.
8. 111U = A<->B->C.
9. 030T = A->B<-C, A->C.
10. 030C = A<-B<-C, A->C.
11. 201 = A<->B<->C.
12. 120D = A<-B->C, A<->C.
13. 120U = A->B<-C, A<->C.
14. 120C = A->B->C, A<->C.
15. 210 = A->B<->C, A<->C.
16. 300 = A<->B<->C, A<->C, complete subgraph.
```

The top panel reports the numbers of directed triads across the four waves that had each of the 16 possible configurations. The lower panel provides a key to the configuration names. A few observations illustrate how the triad census can provide insights. The number of "empty" triads (003) declined from 64,330 to 38,886 over the term (and, most rapidly during the time before the first mid-term, which was between wave 1 and wave 2). 021U and 021D (out-star and in-star) configurations changed in step, reflecting the ideas that activity processes and processes of preferential attachment are equally likely to occur in this network. A relatively large amount of "strong cliques" (triads with all 6 ties present, or three reciprocated relationships) emerged, compared to the numbers of configurations close to these "complete subgraphs".

Observations like these should focus ones attention on two things: the likelihood of different mixes of configurations changes as the overall density of the graph increases; and, one is led to wonder what kinds of actors (i.e. actors with what attributes) are more or less likely to be involved in which kinds of local structures.

Let's simplify the picture now, and take a look at the triad census for the four waves where the data have been symmetrized (*Transform>Symmetrize*) using the "maximum" method (if either A ->B, or B->A, then A<->B.

Figure 7.2. UCINET's Triad Census for Student Acquaintanceship, Symmetric (Maximum Method)



```
Triad Census for dataset allwaves_2011-maxsym

                  1         2         3         4
          wave1_201 wave2_201 wave3_201 wave4_201
          --------- --------- --------- ---------
 1  003       64330     51903     49490     38886
 2  012           0         0         0         0
 3  102        3110     13914     15878     23238
 4 021D           0         0         0         0
 5 021U           0         0         0         0
 6 021C           0         0         0         0
 7 111D           0         0         0         0
 8 111U           0         0         0         0
 9 030T           0         0         0         0
10 030C           0         0         0         0
11  201          80      1591      1982      4839
12 120D           0         0         0         0
13 120U           0         0         0         0
14 120C           0         0         0         0
15  210           0         0         0         0
16  300           5       117       175       562
```

We see that the relative numbers of triads with zero, one, two, and three ties change rather remarkably as the total density of the graph increases. Students are increasingly likely to become embedded in local structures that are more closed and dense. In looking at a triad census like the one in figure 7.2, it makes sense to ask about the ratio of empty triads to triads that are not empty; then to ask about the ratio of triads with two or more ties to those with one tie; then to ask about the ratio of triads with all three ties to those with two. That is, the triad census reflects an inherent underlying hierarchy. You cannot have a structure with three ties until you have a structure with two, etc.

Exponential random graph models predict the presence or absence (or the strength) of ties between pairs of actors. In doing so, they explicitly recognize that the likelihood of a given tie is affected by the overall density of the graph, and the overall tendencies in the graph for local ties to be reciprocated, to form local closed structures, and for the overall degree distribution of the graph to be unequal (preferential attachment). Going beyond these tendencies, ERG models ask what kinds of actors and dyads are more or less likely to have ties, within the constraints imposed by the overall structural biases of the graph.

7.3 P1 Testing for Structure

The earliest stochastic model of graph structure to gain wide-spread use was proposed by Holland and Leinhardt in 1981. The P1 model continues to be a very useful tool for describing the structure of a graph. It also is a good way of understanding some of the basic ideas underlying the exponential random graph models that we will examine in the next chapter.

Suppose that we had a set of *k* nodes, with no ties among them. Let's select two of the nodes at random, and add a tie. Now, let's select two nodes again, including the two nodes that are already connected, and add a tie between these two (unless, by chance, we happened to select the same two nodes that already had a tie). We can continue this process of random graph construction until all pairs are connected (i.e., the density is 1).

As the density of our random graph increases, structure emerges. When we add the first tie, we have created a graph with one dyad and *k* - 2 isolates. This is the only possible emergent structure, so that we know it has a probability of 1.0 in a random graph with only one tie. Another thing to note is that the "degree distribution" has changed as we added a tie. Rather than all nodes having a degree of zero, we now have two nodes with a degree of 1, while all the remaining nodes have a degree of 0.

When we add the second tie, there are two possible emergent structures. Either we form a structure with two dyads and *k* - 4 isolates (the most likely outcome), or we create a structure with one 3-node "line" and *k* - 3 isolates. With two ties, there are only two possible graph structures, and we can work out the probability of finding one or the other, if the process of adding ties is purely random. There are also a couple of possible degree-distributions. If we observe the "3-line" structure, there is one node with a degree of 2, two with degree of 1, and all the rest have degree of zero. If we observe the two-dyad graph, there are four nodes with degree of 1, and all the rest have a degree of zero.

The numbers of basic structures in a graph (dyads, lines, triangles, etc.), and the shape of the degree distribution of the graph, change as the density of the graph increases.

As we add the third tie, there is a new range of possibilities for emergent structures. We could now have three dyads and remaining isolates, or we could have a line of 4 connected nodes and remaining isolates, or we could have a single closed triad and remaining isolates, etc. The range of structures that could emerge as we add ties grows exponentially with the increasing density. One way we can represent the possible graphs is as a "tree" (or first-order Markov process). Indeed, it is even possible to work out the probability of each particular kind of graph of a given density. Each of the possible graphs also has an associated degree distribution. In some graphs, there will be nodes with high degree and there will be considerable inequality in the degree distribution. In other possible graphs, the degree distribution will have few "stars" (actors with high degree), and a more equal distribution.

The important lesson from random graphs is that they display emergent structure (cliques, triads, lines, and unequal "social capital") that can arise entirely by random and path-dependent processes that do not depend on the attributes of the actors. For any given level of density, there is a determinate probability distribution of structural properties of a graph based entirely on random processes.

Real social actors, we suspect, don't build social networks by adding ties at random as just described. But, there are two very interesting and useful things about this "random graph" model.

First, to some degree at least, ties probably do form "at random" in social groups. So, when we see a network that displays "lines" and "dyads" and "closed triads" and "4-lines" we need to be careful to not over-interpret what is going on. In some graphs of a given density, it is possible to observe a single "leader" with very high degree – and most other nodes with very low degree – and this can happen by purely random processes. So, when we look at a real social network, it is very helpful to remember that there is some chance that what we are seeing is, in fact, the realization of a purely random statistical process, rather than non-random social organizing processes.

Second, the random graph model is also the simplest baseline model of how emergent structure may be a "cause" in itself. Sociologists have proposed a number of theories of how social networks form that depend on structure itself, rather than the attributes of individual actors. "Preferential attachment" (Barabási & Albert, 1999) suggests that actors who have more ties (perhaps for entirely random reasons) may be more attractive as network partners and hence garner new partners at a preferential rate as density increases. In directed graphs, we might find a tendency for "reciprocity" so that if a tie already exists from A to B, it may well be that the next tie is more likely to emerge from B to A than any other possibility. If A is already connected to B, and B is already connected to C, the "transitive" theory (Wasserman & Faust, 1994) suggests that an AC tie is more likely to develop next than any other random tie. To judge whether any of these kinds of structural processes are actually present in an observed network, we need to know whether the observed numbers of "stars," "reciprocated ties," or "transitive triads" (for example) differ from what would happen entirely by random processes.

The P1 model is a method of describing the structure of an observed network in terms of some of these basic structural processes. The unit of observation, or case, in P1 (and the other stochastic models of graphs) is the dyad. For a non-directed graph, each dyad can be of one of two types: null (no tie), or present (tie). For a directed graph, each dyad is one of three types: null (no tie), asymmetric (a tie from A to B, or from B to A, that is not reciprocated), or mutual (a reciprocated tie between A and B).

The P1 model hypothesizes that the probability a given dyad is null, asymmetric, or mutual is the realization of four structural processes. The fitted model assigns parametric values to these four processes. The first key parameter is theta (θ), which reflects the effect of total density on the probability that any given dyad is asymmetric or mutual, instead of null. The second key parameter is alpha (α), which reflects the out-degree or "expansiveness" of each node (there are as many alpha parameters as there are nodes). Alpha, then, fits the out-degree distribution of the graph, and reflects individual differences in the propensity to seek ties. The third key parameter is beta (β), which describes the "attractiveness," or "popularity," or "status" of nodes by modeling variation in the in-degree of nodes. Again, there are as many beta parameters as there are nodes. The fourth key parameter is rho (ρ) which reflects the tendency toward reciprocity (that is, given that a dyad contains one tie, what is the probability that it contains a second?). The P1 model does not directly address the question of closure or transitivity. These structural aspects were subsequently addressed with additional P and ERG models.

The parameters of the P1 model are estimated by fitting three simultaneous equations to the dyadic data (From Borgatti, et al. UCINET 6):

$$n_{ij} = \lambda_{ij} \tag{7.1}$$

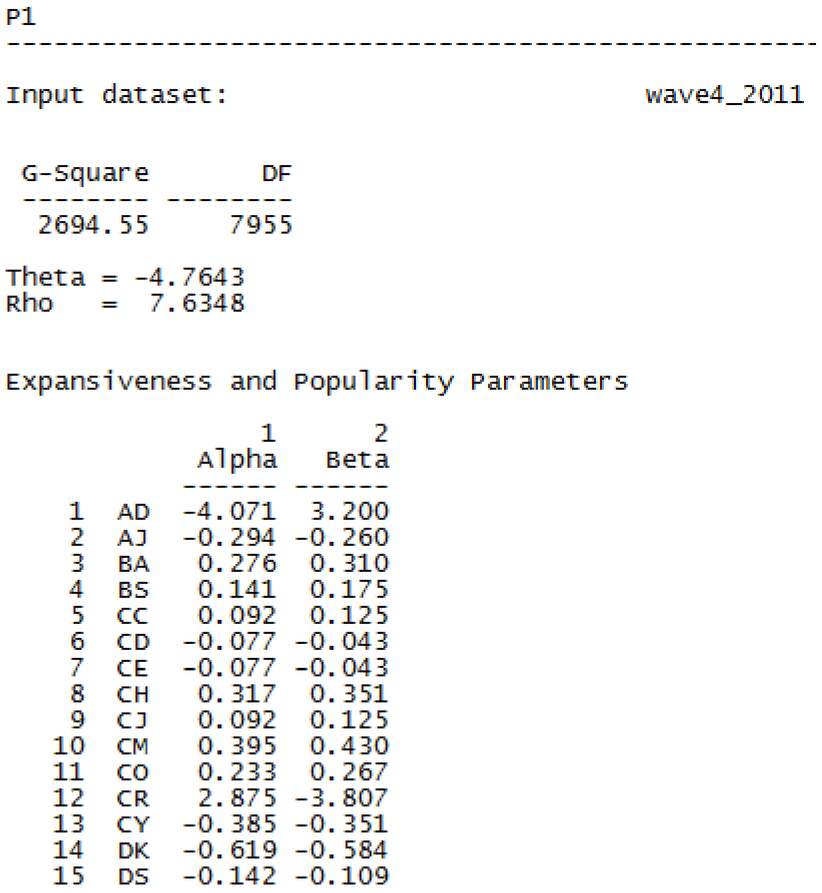
$$a_{ij} = \lambda_{ij} e^{(\theta+\alpha_i+\beta_j)} \tag{7.2}$$

$$m_{ij} = \lambda_{ij} e^{(\rho+2\theta+\alpha_i+\alpha_j+\beta_i+\beta_j)} \tag{7.3}$$

In the equations above, n_{ij} represents the probability that a given tie in a network will be null, a_{ij} represents the probability it will be asymmetric, and m_{ij} represents the probability it will be mutual. The model suggests that the probability a given dyad is null is equal to a constant, lambda. Lambda is simply a scalar to assure that the probabilities of the various types sum to 1.0. The probability that a dyad is asymmetric is a function of the overall graph density plus the expansiveness and attractiveness of the two nodes in the dyad. The likelihood that a dyad is mutual (reciprocated ties) is a function of overall density, the expansiveness of both actors, the attractiveness of both actors, and an additional graph-wide propensity (rho) for reciprocity. Note that the model is log-additive (or multiplicative). That is, the effects of the out-degree distribution, in-degree distribution, and reciprocity multiply modify the effects of density.

Figure 7.3 shows a portion of the results of running *UCINET>Network>P1* on the asymmetric acquaintanceship ties in our social networks class at the end of the academic term (Wave 4).

Figure 7.3. UCINET's P1 Analysis of Wave 4 (Asymmetric) Student Acquaintanceship



```
P1
--------------------------------------------------

Input dataset:                          wave4_2011


 G-Square       DF
 -------- --------
  2694.55     7955

Theta = -4.7643
Rho   =  7.6348


Expansiveness and Popularity Parameters

                  1      2
              Alpha   Beta
             ------ ------
    1  AD   -4.071  3.200
    2  AJ   -0.294 -0.260
    3  BA    0.276  0.310
    4  BS    0.141  0.175
    5  CC    0.092  0.125
    6  CD   -0.077 -0.043
    7  CE   -0.077 -0.043
    8  CH    0.317  0.351
    9  CJ    0.092  0.125
   10  CM    0.395  0.430
   11  CO    0.233  0.267
   12  CR    2.875 -3.807
   13  CY   -0.385 -0.351
   14  DK   -0.619 -0.584
   15  DS   -0.142 -0.109
```

The first bit of information given in the output is the "badness of fit" (G-square) statistic and associated degrees of freedom. As UCINET's documentation suggests, it is difficult to interpret this value precisely because the distributional assumptions are unknown. A rough comparison to the chi-square distribution with the same degrees of freedom suggests that the P1 model leaves significant residual variation (the residuals are available as output).

Next, the estimated parameters of the model are given. Theta (the density parameter) is -4.7643. The exponentiated value is 0.0085. Rho has a large positive value (7.6348), which suggests that if there is a single tie between two actors, the odds that there is a second (reciprocating) tie are over seven times as large as we would expect in a random graph of the same overall density, controlling for the out and in degrees of the members of the dyad. That is, there is a notable tendency toward reciprocation in acquaintanceship nominations which is hardly a surprising result.

Last, the parameter for the expansiveness and attractiveness of each actor are shown. We see, for example, that actor AD tends to initiate fewer ties than we would expect (alpha), but to receive more in-ties than we would expect in a random model (beta). Examining the distribution of these parameters allows us to see the shape of the in-degree distribution and the out-degree distribution. This describes the extent to which our graph displays unequal expansiveness and attractiveness, controlling for overall density and the observed tendency toward reciprocation.

UCINET's implementation of P1 is a purely descriptive tool to identify actors who are in-stars, and out-stars, and to evaluate the magnitude of reciprocity (controlling for total density and the observed degree distributions). To test the statistical significance of the parameters, one must generate a large number of random graphs of the same density, fit the P1 model, and develop sampling distributions of the alpha, beta, and rho parameters. This computationally intensive process is characteristic of the methodology of all stochastic graph models, and is best pursued in software specifically designed for the purpose (e.g. ERGM, PNET, Siena).

The P1 model itself has largely been supplanted by more recent developments (see particularly Harris, 2014 for an excellent history), that allow more complex structural effects, hypothesis testing, and the inclusion of actor attributes and dyadic attributes. The P1 model, though, represents a major step forward in modeling (i.e. predicting and testing hypotheses about) network structures. Among the most important of the lessons of the P1 model are:

> The idea of treating the dyad as the unit of analysis as a way of approaching the prediction of social structure.
>
> The use of the Markov approach to understand how the probabilities of random graphs with different structures emerge with increases in density.
>
> The recognition that the textures of social networks may be generated by structural processes, as well as the agency of actors (e.g. preferential attachment, reciprocity, and closure). And, that these structural processes form a hierarchy of multiplicative (not additive) effects.

7.4 Summary

In this chapter, we introduced many of the key concepts used in models of network selection. We discussed the ways that a network structure itself can become the dependent variable of predictive models. We also introduced network self-organization processes, actor attribute processes, and exogenous context as three mechanisms guiding the evolution of network structures.

Additionally, we outlined the importance of the triad census and how it can be computed in UCINet. We concluded the chapter with an introduction to the P1 model and a discussion of the importance of modeling random network processes.

7.5 References

Barabási, Albert-Laszlo and Reka Albert. 1999. "Emergence of Scaling in Random Networks," *Science*, 286, 509-512.

Harris, Jenine K. 2014. *An Introduction to Exponential Random Graph Modeling*. Los Angeles: Sage Publications (Quantitative Applications in the Social Sciences series #173).

Holland, Paul W., and Samuel Leinhardt. 1981. "An Exponential Family of Probability Distributions for Directed Graphs." *Journal of the American Statistical Association,* 76(373):33-6.

Lusher, Dean and Garry Robins. 2013. "Formation of Social Network Structure." Pp. 16-28 in *Exponential Random Graph Models for Social Networks*, edited by D. Lusher, J. Koskinen, and G. Robbins. Cambridge: Cambridge University Press.

Wasserman, Stanley, and Katherine Faust. 1994. *Social Network Analysis: Methods and Applications*. Cambridge: Cambridge University Press.

Chapter 8. Models of Network Selection - ERGM and Multi-level Models

*** COMING SOON ***

Chapter 9. Models of Network Dynamics and Co-evolution

*** COMING SOON ***